The EMBO Journal Vol.2 No.3 pp.453-462, 1983

The DNA sequence of the H-2K^b gene: evidence for gene conversion as a mechanism for the generation of polymorphism in histocompatibilty antigens

E. Weiss[1,2], L. Golden[1,2], R. Zakut[3], A. Mellor[1,2], K. Fahrner[1,2], S. Kvist[4] and R.A. Flavell[1,2]*

[1]Laboratory for Gene Structure and Expression, National Institute for Medical Research, MRC, Mill Hill, London, UK, [2]Biogen Research Corporation, 241 Binney St., Cambridge, MA 02142, USA, [3]Weizmann Institute of Science, Rehovot, Israel, and [4]European Molecular Biology Laboratory, Postfach 10.2209, Meyerhofstrasse 1, D-6900 Heidelberg, FRG

Communicated by R.A. Flavell
Received on 17 January 1983

We have determined the DNA sequence of the H-2K^b gene of the C57B1/10 mouse. Comparison of this sequence with that of the allelic H-2K^d shows surprisingly that the exons have accumulated more mutations than their introns. Moreover, many of these changes in the exons are clustered in short regions or hot spots. Additional comparison of these sequences with the H-2L^d and H-2D^b sequences shows that, in several cases, the altered sequence generated at the hot spot is identical to the corresponding region of a non-allelic H-2 gene. The clustered changes are responsible for 60% of the amino acid differences between the H-2K^b and H-2K^d genes and suggest that micro-gene conversion events occurring within the exons and involving only tens of nucleotides are an important mechanism for the generation of polymorphic differences between natural H-2 alleles.

Key words: H-2 complex/gene conversion/H-2 K^b gene/polymorphism/histocompatibility antigens

Introduction

The class I genes of the major histocompatibility complex (MHC) encodes a 40 000 - 50 000 polypeptide (called H-2 in the mouse and HLA in man) which is an integral membrane protein associated with β2-microglobulin. DNA and protein sequence information shows that class I genes consist of eight exons encoding a leader sequence, three extracellular globular domains (α_1, α_2 and α_3), a trans-membrane segment and three cytoplasmic domains (Steinmetz *et al.*, 1981).

The major class I H-2 genes are H-2K, H-2D and H-2L found closely linked to each other and other H-2 related genes such as Qa and TL on chromosome 17 of the mouse (Figure 1). H-2 genes are highly polymorphic, and ~50 alleles at each of the H-2K and H-2D loci have been observed (see Klein, 1979). The polymorphic differences are greatest in the α_1 and α_2 extracellular domains. The protein sequences of two non-allelic gene products in a given H-2 haplotype are in some cases more similar to one another (e.g., H-2K^b and H-2D^b) than to their respective alleles. This suggested to us (Flavell *et al.*, 1982a) and others (Lalanne *et al.*, 1982; Evans *et al.*, 1982) that the polymorphic differences may result, at least in part, by gene conversion-like events which exchange genetic information between non-alleles by a copying mechanism.

To test this we have sequenced the H-2K^b gene and compared this with its allelic H-2K^d gene sequence and the sequence of the non-allelic H-2D^b and H-2L^d genes. We have shown previously (Flavell *et al.*, 1982b; Weiss, *et al.*, 1983) that a mutant H-2K^b gene, called H-2K^{bml}, results from a micro-gene conversion-like event which causes the introduction of a new stretch of DNA sequence into the H-2K^b gene over a short DNA segment of the order of 30 nucleotides. This new sequence is identical to that of the H-2L^d gene in the same region which suggests that an H-2L^d-like gene may have donated this new information. Pease *et al.* (1983) have made a similar suggestion by comparing the amino acid sequence of mutant and normal H-2 genes. In this paper we show that several of the differences in nucleotide sequence between the H-2K^b and H-2K^d alleles can be explained by similar micro-gene conversions. We therefore suggest that this mechanism plays an important role in the generation of polymorphism in H-2 genes in the mouse population.

Results

We have described previously the cloning of the H-2K^b gene from a cosmid library made with DNA from the B/10 mouse (Mellor *et al.*, 1982). The sequence of this gene was determined by the Maxam and Gilbert (1980) and M13 procedures; the strategy used is shown in Figure 2. The DNA sequence is shown in Figure 3 and is complete except for an ~1200 nucleotide gap in the large intron between exons 3 and 4.

General structure of the H-2K^b gene

The overall structure and dimensions of the H-2K^b gene is the same as that of other H-2 genes sequenced. The H-2K^b gene consists of eight exons and seven introns. Table I shows that the majority of the exons are of the same length as those of the H-2K^d and H-2L^d genes. However, the leader exons of the H-2K^b (this paper) and H-2K^d (Kvist *et al.*, 1983) gene encode one more amino acid than the leader of the H-2L^d gene; and the trans-membrane exon of the H-2K^b gene encodes one more amino acid than the corresponding exon of H-2K^d (lacks residue 312) and H-2L^d (lacks residue 309).

The approximate size of the introns is the same in all three genes with minor variation in the exact lengths (Table I). The large intron between exons 3 and 4 is of a similar length in the H-2K alleles but is considerably shorter in the H-2L^d gene and

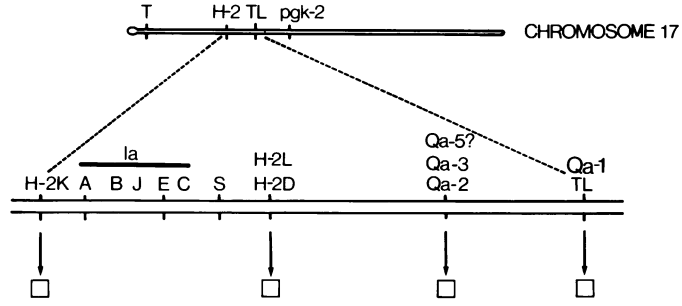


Fig. 1. Genetic map of the murine H-2 and associated loci. The top line shows a diagram of chromosome 17 with the centromere on the left. The bottom line shows an expanded diagram of the regions between the H-2K locus on the left and the TL locus on the right. H-2 class I molecules (□) are expressed from at least four separate loci as shown by the arrows below the line.

*To whom reprint requests should be sent.

longer in the H-2-like gene found in the Qa2,3 region (Steinmetz *et al.*, 1981).

The sequence of a cloned H-2K^b cDNA, derived from C56B1/6 (= B6) mice, has been reported recently (Reyes *et al.*, 1982a); the sequence spans the region encoding residues 65 to the 3′-untranslated region (and hence exons 2–8). There are three nucleotide differences between the cDNA sequence and our H-2K^b gene sequence from the B10 mouse. We have also sequenced most of the mutant H-2K^{bml} gene which is also derived from the B6 mouse and in one of these nucleotides (codon 223 in exon 4) the bml and B10 sequences are the same (GTC) and differ from the cDNA sequence (GTT). The other two nucleotides are in the 3′-untranslated region which has not been sequenced in the H-2K^{bml} gene. We have, however, found a single polymorphic difference between the H-2K^{bml} and H-2K^b gene sequences in intron 1; this may be a B6/B10 polymorphism.

In this article we compare the H-2K^b gene sequence with the H-2K^d, L^d genes and the Qa-linked H-2 gene (called the 27.1 gene) of Steinmetz *et al.* (1981) and the 'A' gene of Mellor *et al.* (1982) which is closely linked to the H-2K^b gene.

5′-flanking sequences

The 5′ end of the H-2K^b mRNA has not yet been determined so we cannot define with certainty the 5′ boundary of the gene. Inspection of the sequence of the H-2K^b gene, however, identifies the TATAAA and CCAAT sequences that have been previously found in numerous other eukaryotic genes and which in those cases have been shown to be at least part of the promoter for RNA synthesis by RNA polymerase II. The position of these sequences relative to one another and to the AUG initiation codon is unusual in the H-2K^b gene. The TATA sequence is found 55 nucleotides upstream from the AUG; since RNA polymerae II initiates RNA synthesis ~30 nucleotides downstream from the 5′ nucleotide of the TATA (see e.g., Grosveld *et al.*, 1982) it would follow that this would generate a 5′-untranslated regin of only ~25 nucleotides on H-2K^b mRNA which is short. There is a CCAAT sequence present a further 22 nucleotides 5′ to the TATA sequence. This is unusually close to the TATA, since normally the spacing of the TATA and CCAAT sequences is ~40 nucleotides (see, e.g., Efstratiadis *et al.*, 1980). This spacing is however only 32 nucleotides in the case of the human δ-globin gene. Two other sequences which are homologous to CCAAT sequences are seen at −84 to −80 (CCATT) and −64 to −60 (CCAAG).

A third component of the RNA polymerase II promoter has been identified recently (MacKnight *et al.*, 1981; Grosveld *et al.*, 1982; P. Diercks and C. Weissmann, personal communication). In the case of the rabbit β-globin gene three repeats of a sequence ACCC have been identified (Grosveld *et al.*, 1982; P. Diercks and C. Weissmann, personal communication). The sequence ACCCC is present in the H-2K^b gene at a site −84 to −88 upstream from a putative cap site deduced from position of the TATA box. This may therefore correspond to the '−90' promoter element for the H-2K gene.

The 5′-flanking regions of the two alleles H-2K^b/K^d are very similar as would be expected for true alleles. There is a total of seven nucleotide differences in this region over the 176 base pairs that we have sequenced in the 5′-flanking DNA of the H-2K^b gene (Figure 4; Table I). In both H-2K alleles and the 27.1 gene, the TATA (for the 27.1 gene TGTA) CCAAT sequences are in the same position and no other such sequences can be detected in the 5′-flanking DNA, which in the case of the Qa linked H-2 gene sequence extends to >900 nucleotides upstream from the leader sequence. We think it likely, therefore, that these conserved sequences constitute the H-2K promoter.

The 5′-flanking sequence of the H-2L^d sequence is totally non-homologous to that of the H-2K genes over the 110 nucleotides upstream from the leader sequence available for comparison and no TATA or CCAAT sequences can be identified in this region of the H-2L^d gene.

Sequence conservation of intronic DNA in the H-2K^b gene

Comparision of the intron sequences of the H-2K^b and H-2K^d genes shows 3–4% divergence, similar to the level seen in the 5′-flanking DNA. Introns 1–3 of the non-alleles H-2K^b and H-2L^d show ~10% sequence divergence but introns 4 and 5 are 5–6% divergent (Table I). Introns 6 and 7 are somewhat more conserved in showing 2.3% and 1.8% differences, respectively (Table I). The small difference in the length of the introns results from short deletions/insertions of up to seven nucleotides. The remaining differences result from base substitutions scattered through the introns except

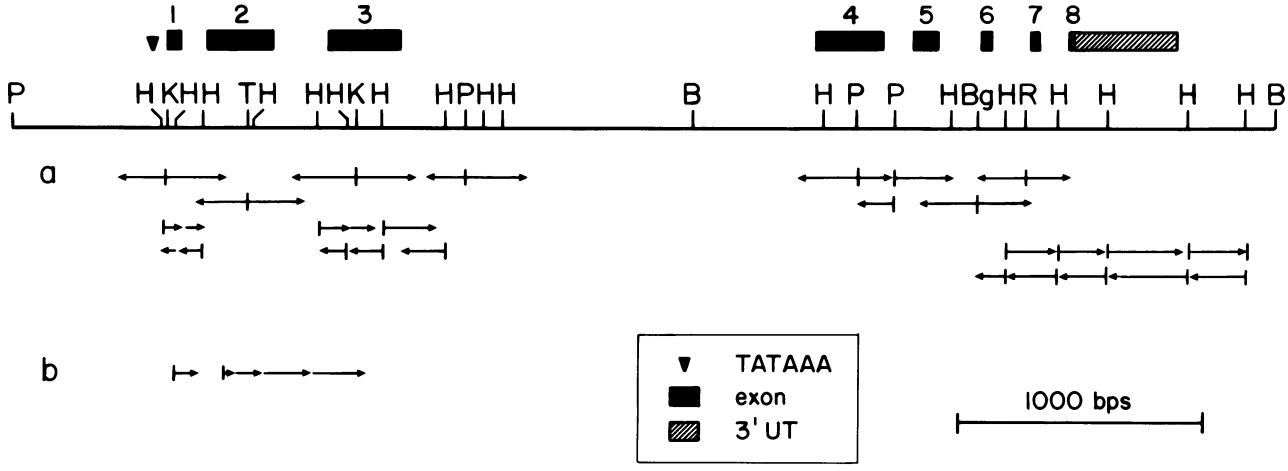


Fig. 2. Restriction map of the H-2K^b gene and sequence strategy. B = *Bam*HI; Bg = *Bgl*II; H = *Hinf*I; K = *Kpn*I; P = *Pst*I; R = *Rsa*I; T = *Taq*I. The *Hinf*I restriction sites are indicated for the sequenced part of the gene. Only the *Taq*I and *Rsa*I site used for sequencing are shown. (**a**) Maxam-Gilbert; (**b**) M13-cloning strategy. The straight arrows give the direction and length of the fragment sequenced. Above the restiction map is shown the position of the TATAAA box, the exons and the 3′-untranslated region.

Fig. 3. Nucleotide sequence of the H-2K^b gene. The amino acid translation of each exon is given below the nucleotide sequence. XXX indicate the gap of ~1200 nucleotides in intron 3. The CCAAT, TATAAA, TGA stop codon and AATAAA sequences are underlined.

in one case, in intron 1, where four out of seven consecutive nucleotides are altered.

The percentage G + C of introns 1 and 2 is remarkably high, 77% and 70%, respectively. In the 50 nucleotides adjoining exons at the 3′ ends of these introns, the G + C content exceeds 90%. The remaining introns are from 40–58% G + C (Table I).

Sequence divergence in H-2 exons

Previous comparisons of the sequences of a given gene (such as a globin gene) have shown that the intronic DNA sequences show more divergence than mRNA-coding sequences (van den Berg *et al.*, 1978; Efstratiadis *et al.*, 1980).

This generality does not, however, hold for the pair of H-2K^b and K^d alleles. The apparent extent of sequence divergence of the exons is ~3-fold greater than that of the introns for exons 2 and 3 which encode the polymorphic extracellular α_1 and α_2 domains of the H-2 heavy chain, and 2-fold greater than that of the introns for the leader sequences, α_3 and trans-membrane domains.

Careful inspection of the distribution of nucleotide sequence changes shows that these are scattered in the introns as stated above, but consist of both scattered changes and a number of clusters containing numerous changes in the exons. If we subtract the clustered changes, then the frequency of nucleotide substitutions in the introns and exons becomes similar (Table I). This suggests further that the scattered changes result from a similar mechanism in both introns and exons, and, conversely, that a different mechanism is responsible for the clustered sequence changes. We shall consider below the nature of these clustered sequence differences between the alleles.

As discussed above, the extent of sequence divergence for the non-coding regions of the H-2K^b and H-2L^d genes is greater than the corresponding sequences of the alleles. In contrast, the sequence divergence for the exons is similar; that is, the divergence of the two non-alleles H-2K^b/H-2L^d is about the same as the two alleles H-2K^b/H-2K^d (Table I).

A mosaic pattern of DNA sequence in exons of H-2 genes

As already discussed, we have noted that sequence differences in the exons of the H-2K^b and K^d genes tend to be clustered in hot spots. We have therefore analysed the changes in these clusters to look for a potential mechanism. We have noted one such clustered difference in the leader sequence, three in the α_1 domain (exon 2), five such clusters in

Table I

	Length [bps]			Different nucleotides*		Changes K^b/K^d		% changes*			% G + C		
	K^d	K^d	L^d	K^b/K^d	K^b/L^d	Coding	Silent	Non-clustered K^b/K^d	Total K^b/K^d	Total K^b/L^d	K^b	K^d	L^d
5′ region[a]	[176	241	110]	7	b			0	3.9	b	59	60	62
Leader exon 1	64	64	61	5	7	1	4	3.1	7.8	10.9	72	73	77
intron 1	190	197	178[c]	14	46				3.7	24	77	77	74
α1 exon 2	270	270	270	28	30	22	6	4.0	10.3	11.1	65	64	65
intron 2	186	187	185	8	25				4.3	13.5	70	70	70
α2 exon 3	276	276	276	31	35	24	7	3.2	11.2	12.7	62	64	62
intron 3	1700[d]	1729	1100[d]	19[d]	46[d]				4.0[d]	9.7[d]	56[g]	49	49[g]
α3 exon 4	276	276	276	17	11	14	3	3.2	6.1	4.0	58	57	58
intron 4	127	127	128	6	7				4.7	5.5	58	58	58
TM exon 5	120	117	117	9	18	8	1	1.7	5.8	13.3	52	49	51
intron 5	178	178	177	7	9				3.9	5.0	50	50	46
exon 6	33	33	33	2	0	2	0	6.0	6.0	0	58	52	56
intron 6	172	173	173	4	8				2.3	4.6	43	43	44
exon 7	39	39	39	1	4	1	0	2.6	2.6	10.2	51	54	49
intron 7	111	112	112[e]	2	10				1.8	9.0	52	51	53
exon 8	32	32	32[e]	0	1	0	0	0	0	3.1	47	47	44
3′ UT	424	414	[310][f]	23	36				5.0	11.6	45	46	52
3′-flanking[a]	274	676		15					3.0		54	50	

*Total different nucleotides counts each difference irrespective of the event, i.e., a deletion/insertion of seven nucleotides counts as 7. In the % changes we count a deletion/insertion as a single event irrespective of its length.

[a]The lengths refer to the DNA sequence available.

[b]These sequences are no more similar than random.

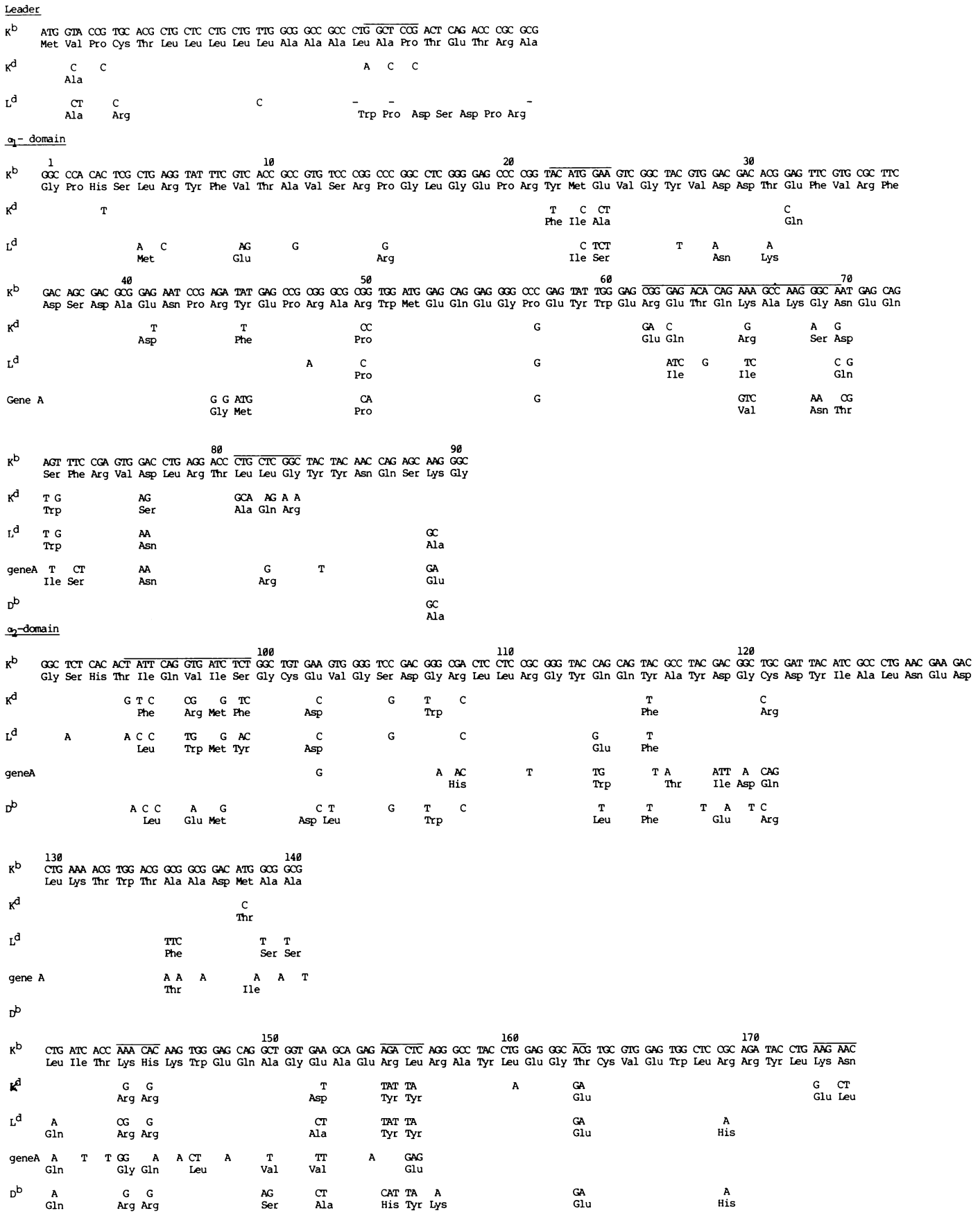
[c]We have used the sequence of Evans *et al.* (1982) for the first intron as this sequence is complete. The sequence of Moore *et al.* (1982) is used for the remainder of the gene because it is more extensive.

[d]The length of intron 3 of the K^b and L^d sequences are approximate and deduced from extensive restriction site mapping. We have determined 471 nucleotides of intron 3 of the K^b gene and the % differences refer to this region only.

[e]Exon 8 of Moore *et al.* (1982) is only depicted as five nucleotides. Exon 8 of H-2K^b can be unambiguously determined to be 32 nucleotides because the H-2K^b sequence is determined. Since these 32 nucleotides are homologous in the H-2L^d gene, we cite this alternative possibility for the H-2L^d which would extend the predicted length of the H-2L^d protein in Moore *et al.* (1982).

[f]The 3′-untranslated region of the H-2K^b gene is only homologous to the H-2L^d gene for 310 nucleotides. After this, the two sequences are totally different. The % homology is calculated only over these 310 nucleotides.

[g]The K^b and L^d sequences are for only part of the intron 3 whereas the K^d sequence is for the entire intron.


Leader
Kb ATG GTA CCG TGC ACG CTG CTC CTG CTG TTG GCG GCC GCC CTG GCT CCG ACT CAG ACC CGC GCG
Met Val Pro Cys Thr Leu Leu Leu Leu Leu Ala Ala Ala Leu Ala Pro Thr Glu Thr Arg Ala
Kd C C A C C
Ala
Ld CT C C - - -
Ala Arg Trp Pro Asp Ser Asp Pro Arg
α1- domain
1 10 20 30
Kb GGC CCA CAC TCG CTG AGG TAT TTC GTC ACC GCC GTG TCC CGG CCC GGC CTC GGG GAG CCC CGG TAC ATG GAA GTC GGC TAC GTG GAC GAC ACG GAG TTC GTG CGC TTC
Gly Pro His Ser Leu Arg Tyr Phe Val Thr Ala Val Ser Arg Pro Gly Leu Gly Glu Pro Arg Tyr Met Glu Val Gly Tyr Val Asp Asp Thr Glu Phe Val Arg Phe
Kd T T C CT C
Phe Ile Ala Gln
Ld A C AG G G C TCT T A A
Met Glu Arg Ile Ser Asn Lys
40 50 60 70
Kb GAC AGC GAC GCG GAG AAT CCG AGA TAT GAG CCG CGG GCG CGG TGG ATG GAG CAG GAG GGG CCC GAG TAT TGG GAG CGG GAG ACA CAG AAA GCC AAG GGC AAT GAG CAG
Asp Ser Asp Ala Glu Asn Pro Arg Tyr Glu Pro Arg Ala Arg Trp Met Glu Gln Glu Gly Pro Glu Tyr Trp Glu Arg Glu Thr Gln Lys Ala Lys Gly Asn Glu Gln
Kd T T CC G GA C G A G
Asp Phe Pro Glu Gln Arg Ser Asp
Ld A C G ATC G TC C G
Pro Ile Ile Gln
Gene A G G ATG CA G GTC AA CG
Gly Met Pro Val Asn Thr
80 90
Kb AGT TTC CGA GTG GAC CTG AGG ACC CTG CTC GGC TAC TAC AAC CAG AGC AAG GGC
Ser Phe Arg Val Asp Leu Arg Thr Leu Leu Gly Tyr Tyr Asn Gln Ser Lys Gly
Kd T G AG GCA AG A A
Trp Ser Ala Gln Arg
Ld T G AA GC
Trp Asn Ala
geneA T CT AA G T GA
Ile Ser Asn Arg Glu
Db GC
Ala
α2-domain
100 110 120
Kb GGC TCT CAC ACT ATT CAG GTG ATC TCT GGC TGT GAA GTG GGG TCC GAC GGG CGA CTC CTC CGC GGG TAC CAG CAG TAC GCC TAC GAC GGC TGC GAT TAC ATC GCC CTG AAC GAA GAC
Gly Ser His Thr Ile Gln Val Ile Ser Gly Cys Glu Val Gly Ser Asp Gly Arg Leu Leu Arg Gly Tyr Gln Gln Tyr Ala Tyr Asp Gly Cys Asp Tyr Ile Ala Leu Asn Glu Asp
Kd G T C CG G TC C G T C T C
Phe Arg Met Phe Asp Trp Phe Arg
Ld A A C C TG G AC C G C G T
Leu Trp Met Tyr Asp Glu Phe
geneA G A AC T TG T A ATT A CAG
His Trp Thr Ile Asp Gln
Db A C C A G C T G T C T T A T C
Leu Glu Met Asp Leu Trp Leu Phe Glu Arg
130 140
Kb CTG AAA ACG TGG ACG GCG GCG GAC ATG GCG GCG
Leu Lys Thr Trp Thr Ala Ala Asp Met Ala Ala
Kd C
Thr
Ld TTC T T
Phe Ser Ser
gene A A A A A A T
Thr Ile
Db
150 160 170
Kb CTG ATC ACC AAA CAC AAG TGG GAG CAG GCT GGT GAA GCA GAG AGA CTC AGG GCC TAC CTG GAG GGC ACG TGC GTG GAG TGG CTC CGC AGA TAC CTG AAG AAC
Leu Ile Thr Lys His Lys Trp Glu Gln Ala Gly Glu Ala Glu Arg Leu Arg Ala Tyr Leu Glu Gly Thr Cys Val Glu Trp Leu Arg Arg Tyr Leu Lys Asn
Kd G G T TAT TA A GA G CT
Arg Arg Asp Tyr Tyr Glu Glu Leu
Ld A CG G CT TAT TA GA A
Gln Arg Arg Ala Tyr Tyr Glu His
geneA A T T GG A A CT A T TT A GAG
Gln Gly Gln Leu Val Val Glu
Db A G G AG CT CAT TA A GA A
Gln Arg Arg Ser Ala His Tyr Lys Glu His

```
                                  180
K^b  GGG AAC GCG ACG CTG CTG CGC ACA
     Gly Asn Ala Thr Leu Leu Arg Thr

K^d        T A
            Glu

L^d        T   T

D^b
```

α_3- domain

```
     183                          190                                  200
K^b  GAT TCC CCA AAG GCC CAT GTG ACC CAT CAC AGC AGA CCT GAA GAT AAA GTC ACC CTG AGG TGC TGG GCC CTG GGC TTC
     Asp Ser Pro Lys Ala His Val Thr His His Ser Arg Pro Glu Asp Lys Val Thr Leu Arg Cys Trp Ala Leu Gly Phe

K^d                                 T       CC      T   C    T G T
                                    Tyr     Pro     Ser Gln Val Asp

L^d                   A                     CC      T   A    G  G
                                            Pro     Ser Lys Gly Glu

D^b                   A                     CC      T   A    G  G
                                            Pro     Ser Lys Gly Glu

        210                                 220                                 230                                 240
K^b  TAC CCT GCT GAC ATC ACC CTG ACC TGG CAG TTG AAT GGG GAG GAG CTG ATC CAG GAC ATG GAG CTT GTG GAG ACC AGG CCT GCA GGG GAT GGA ACC TTC CAG
     Tyr Pro Ala Asp Ile Thr Leu Thr Trp Gln Leu Asn Gly Glu Glu Leu Ile Gln Asp Met Glu Leu Val Glu Thr Arg Pro Ala Gly Asp Gly Thr Phe Gln

K^d                 T                                       C     C                       A
                                                          Asp     Thr

L^d                                                             C
                                                                Thr

D^b                                                             C
                                                                Thr

                                    250                                 260                                 270
K^b  AAG TGG GCA TCT GTG GTG GTG CCT CTT GGG AAG GAG CAG TAT TAC ACA TGC CAT GTG TAC CAT CAG GGG CTG CCT GAG CCC CTC ACC CTG AGA TGG
     Lys Trp Ala Ser Val Val Val Pro Leu Gly Lys Glu Gln Tyr Tyr Thr Cys His Val Tyr His Gln Gly Leu Pro Glu Pro Leu Thr Leu Arg Trp

K^d               G                                     A                   C      A                   T
                  Ala                                   Asn                 His    Lys

L^d                                                     A             G     G
                                                        Asn           Arg   Glu

D^b                                                     A             G     G
                                                        Asn           Arg   Glu
```

Transmembrane-domain

```
     275                 280                                 290                                 300
K^b  GAG CCT CCT CCA TCC ACT GTC TCC AAC ATG GCG ACC GTT GCT GTT CTG GTT GTC CTT GGA GCT GCA ATA GTC ACT GGA GCT GTG GTG GCT TTT GTG
     Glu Pro Pro Pro Ser Thr Val Ser Asn Met Ala Thr Val Ala Val Leu Val Val Leu Gly Ala Ala Ile Val Thr Gly Ala Val Val Ala Phe Val

K^d      T                           C   TA  T  A
         Leu                         Thr Val Ile Ile

L^d                G          A   T T      T  T              G              ATG GCC A   T
                              Asp Ser Tyr    Val Ile          Gly            Met Ala Ile Ile

D^b                G          A   T T      T  T              G              ATG GCC A   T
                              Asp Ser Tyr    Val Ile          Gly            Met Ala Ile Ile
```

Exon 6

```
     315                 320
K^b  GGT GGA AAA GGA GGG GAC TAT GCT CTG GCT CCA
     Gly Gly Lys Gly Gly Asp Tyr Ala Leu Ala Pro

K^d                T  A
                  Val Asn

L^d

D^b
```

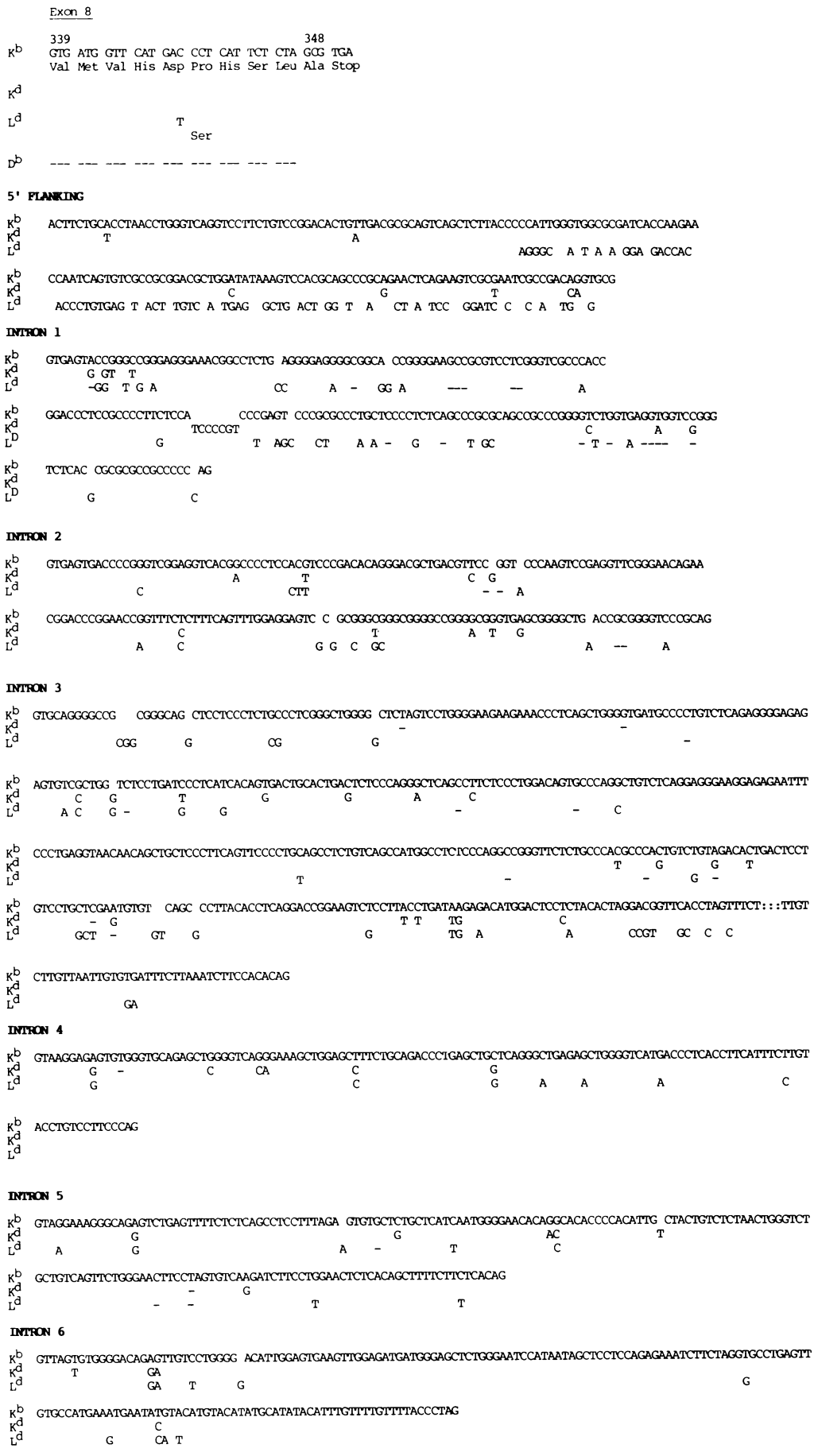
Exon 7

```
     326                 330
K^b  GGC TCC CAG ACC TCT GAT CTG TCT CTC CCA GAT TGT AAA
     Gly Ser Glu Thr Ser Asp Leu Ser Leu Pro Asp Cys Lys

K^d                                      G
                                         Gly

L^d                G       A A          G
                  Ser     Glu Met       Arg

D^b                G       A A          G
                  Ser     Glu Met       Arg
```


Exon 8
339
348
Kb GTG ATG GTT CAT GAC CCT CAT TCT CTA GCG TGA
Val Met Val His Asp Pro His Ser Leu Ala Stop
Kd
Ld T
Ser
Db --- --- --- --- --- --- --- --- ---
5' FLANKING
Kb ACTTCTGCACCTAACCTGGGTCAGGTCCTTCTGTCCGGACACTGTTGACGCGCAGTCAGCTCTTACCCCCATTGGGTGGCGCGATCACCAAGAA
Kd T A
Ld AGGGC A T A A GGA GACCAC
Kb CCAATCAGTGTCGCCGCGGACGCTGGATATAAAGTCCACGCAGCCCGCAGAACTCAGAAGTCGCGAATCGCCGACAGGTGCG
Kd C G T CA
Ld ACCCTGTGAG T ACT TGTC A TGAG GCTG ACT GG T A CT A TCC GGATC C C A TG G
INTRON 1
Kb GTGAGTACCGGGCCGGGAGGGAAACGGCCTCTG AGGGGAGGGGCGGCA CCGGGGAAGCCGCGTCCTCGGGTCGCCCACC
Kd G GT T
Ld -GG T G A CC A - GG A --- — A
Kb GGACCCTCCGCCCCTTCTCCA CCCGAGT CCCGCGCCCTGCTCCCCTCTCAGCCCGCGCAGCCGCCCGGGGTCTGGTGAGGTGGTCCGGG
Kd TCCCCGT C A G
LD G T AGC CT A A - G - T GC - T - A ---- -
Kb TCTCAC CGCGCGCCGCCCCC AG
Kd
LD G C
INTRON 2
Kb GTGAGTGACCCCGGGTCGGAGGTCACGGCCCCTCCACGTCCCGACACAGGGACGCTGACGTTCC GGT CCCAAGTCCGAGGTTCGGGAACAGAA
Kd A T C G
Ld C CTT - - A
Kb CGGACCCGGAACCGGTTTCTCTTTCAGTTTGGAGGAGTC C GCGGGCGGGCGGGGCCGGGGCGGGTGAGCGGGGCTG ACCGCGGGGTCCCGCAG
Kd C T A T G
Ld A C G G C GC A -- A
INTRON 3
Kb GTGCAGGGGCCG CGGGCAG CTCCTCCCTCTGCCCTCGGGCTGGGG CTCTAGTCCTGGGGAAGAAGAAACCCTCAGCTGGGGTGATGCCCCTGTCTCAGAGGGGAGAG
Kd - -
Ld CGG G CG G -
Kb AGTGTCGCTGG TCTCCTGATCCCTCATCACAGTGACTGCACTGACTCTCCCAGGGCTCAGCCTTCTCCCTGGACAGTGCCCAGGCTGTCTCAGGAGGGAAGGAGAGAATTT
Kd C G T G G A C
Ld A C G - G G - - C
Kb CCCTGAGGTAACAACAGCTGCTCCCTTCAGTTCCCCTGCAGCCTCTGTCAGCCATGGCCTCTCCCAGGCCGGGTTCTCTGCCCACGCCCACTGTCTGTAGACACTGACTCCT
Kd T G G T
Ld T - - G -
Kb GTCCTGCTCGAATGTGT CAGC CCTTACACCTCAGGACCGGAAGTCTCCTTACCTGATAAGAGACATGGACTCCTCTACACTAGGACGGTTCACCTAGTTTCT:::TTGT
Kd - G T T TG C
Ld GCT - GT G G TG A A CCGT GC C C
Kb CTTGTTAATTGTGTGATTTCTTAAATCTTCCACACAG
Kd
Ld GA
INTRON 4
Kb GTAAGGAGAGTGTGGGTGCAGAGCTGGGGTCAGGGAAAGCTGGAGCTTTCTGCAGACCCTGAGCTGCTCAGGGCTGAGAGCTGGGGTCATGACCCTCACCTTCATTTCTTGT
Kd G - C CA C G
Ld G C G A A A C
Kb ACCTGTCCTTCCCAG
Kd
Ld
INTRON 5
Kb GTAGGAAAGGGCAGAGTCTGAGTTTTCTCTCAGCCTCCTTTAGA GTGTGCTCTGCTCATCAATGGGGAACACAGGCACACCCCACATTG CTACTGTCTCTAACTGGGTCT
Kd G G AC T
Ld A G A - T C
Kb GCTGTCAGTTCTGGGAACTTCCTAGTGTCAAGATCTTCCTGGAACTCTCACAGCTTTTCTTCTCACAG
Kd - G
Ld - - T T
INTRON 6
Kb GTTAGTGTGGGGACAGAGTTGTCCTGGGG ACATTGGAGTGAAGTTGGAGATGATGGGAGCTCTGGGAATCCATAATAGCTCCTCCAGAGAAATCTTCTAGGTGCCTGAGTT
Kd T GA
Ld GA T G G
Kb GTGCCATGAAATGAATATGTACATGTACATATGCATATACATTTGTTTTGTTTTACCCTAG
Kd C
Ld G CA T

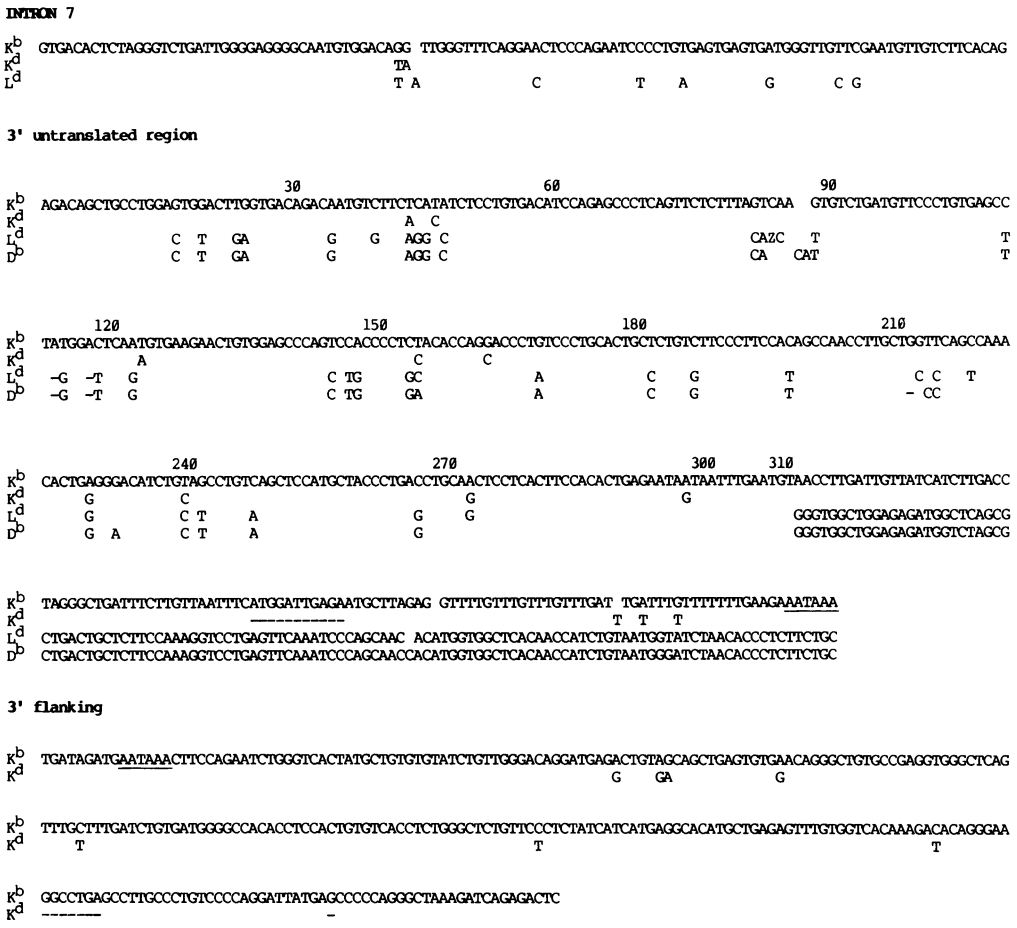


Fig. 4. A comparison of the sequence of the H-2K^b, K^d, L^d and D^b genes. The H-2D^b gene sequence is only available from codon 81 to the end of the mRNA. We also include a comparison with the H-2-like gene closely linked to the H-2K^b gene and called gene A in Mellor *et al.* (1982). This sequence is available from codons 44 to 90 and 99 to 160. Deletions are indicated by dashes; the sequences are identical unless indicated. The clustered changes discussed in the text are overlined.

the α_2 domain, one cluster in the α_3 domain and one in the trans-membrane domain. We shall consider these in order.

Leader sequence. The leader exon of the H-2K^b and H-2K^d genes encodes 21 amino acids; there are five differences in nucleotide sequences and three of these are clustered in a seven nucleotide stretch. The H-2L^d leader is shorter by one codon and alignment of the DNA sequence suggests that this occurred by three separate nucleotide deletions which shift the reading frame and hence the amino acids encoded (Figure 4).

Exon 2; α_1 domain. The second exons of the two H-2K alleles differ at a total of 28 nucleotides. Of these, 17 are clustered as follows. (a) Between the codons for residues 22 – 24. Four of a total of eight nucleotides differ between the two alleles (Figure 4). The sequence in this region in the H-2K^b and H-2K^d genes is not represented in the H-2L^d, H-2D^b (Reyes *et al.*, 1982b) or 27.1 gene. (b) Between the codons for amino acids 62 – 70, six nucleotide differences exist between the H-2K^b and H-2K^d alleles which cause a total of five amino acid substitutions. The entire protein sequence of the H-2K^b and H-2D^b genes is identical from residues 32 to 70 with the exception of one amino acid; the latter difference could result from a single base substitution ($CGC \underset{50}{\rightarrow} CCG$).

The extraordinary extent of homology between the two non-allelic H-2K^b and H-2D^b genes could be the result of a gene conversion event between these two genes. The nucleotide sequence is not available for the H-2D^b gene in this region so a rigorous test of this idea is not yet possible. (c) Seven of a total of nine residues in the codons for amino acids 81 – 83 differ between the H-2K^b and H-2K^d genes. The sequence of the H-2D^b and H-2L^d genes are identical to the H-2K^b gene in this region. The H-2K^b gene may therefore have been converted in this region by the H-2D^b or L^d-like gene. In fact, the H-2K^b and H-2D^b genes are identical from codons 81 to 88.

Exon 3; α_2 domain. Clustered changes are seen at the following sites. (a) Codons 94 – 99, eight of 16 nucleotides differ between the two H-2K alleles. In addition, a further four out of 19 nucleotides differ in the closely linked codons for residues 102 – 108. We have examined all known H-2-like gene sequences available to us (H-2D^b, L^d, H-2-like gene in Qa region and unidentified H-2^d cDNAs in Lalanne *et al.*, 1982), but have not found a sequence homologous to either the H-2K^b and H-2K^d genes in this region. (b) A short cluster is possibly present at codons 144 – 145 with two out of six nucleotides differing between the two allelic H-2K genes. Although this difference is not large, it is notable that the H-2K^b sequence is identical to that of the H-2D^b (Figure 4) and the 27.1 gene (not shown). (c) Codons 155 – 156, where five out of a total of six nucleotides differ between the H-2K^b and H-2K^d genes. The sequence of the H-2K^d gene is identical at this position with the H-2L^d gene. We believe, therefore, that an H-2L^d-like gene has converted the acceptor H-2K^d gene. An additional difference exists between the two H-2K

genes at residue 152; we have not located an H-2 gene which shares the H-2K^b DNA sequence from residues 152 to 156. It should be noted that the H-2K^{bml} mutant gene probably results from a similar gene conversion between a donor H-2L^d-like gene and the H-2K^b gene at essentially the same site (Weiss *et al.*, 1983). (d) The H-2K^d and K^b genes differ at residue 163 (K^d GAG: K^b ACG). The H-2K^d codon is the same as that of the H-2L^d (Figure 4) and 27.1 genes (not shown). (e) The H-2K^b and K^d genes differ from residues 173 to 177 in five out of a total of 14 nucleotides. The H-2K^b sequence is identical to the H-2D^b gene at this position (Figure 4). This change could result from conversion of the H-2K^b gene by an H-2D^b-like donor gene.

Exon 4; α_3-domain. Eight of a total of 24 nucleotides differ between the H-2K^b/K^d genes from residues 191 to 198. Although these changes are obviously clustered, we have not yet found a possible H-2 donor gene among the sequences available to explain these changes.

Exon 5; trans-membrane domain. A single cluster is present between residues 284 and 287 where five out of nine nucleotides differ between the H-2K^b and K^d genes. the H-2K^b gene seqence is identical to that of the 27.1 gene except for one nucleotide (ATG GCG ACC ATT). These allelic differences may result from a conversion between the H-2K^b gene and a gene similar in sequence to the 27.1 gene of Steinmetz *et al.* (1981).

3'-Untranslated region and extragenic DNA sequences. The 3'-untranslated region of the H-2K^b gene extends from the termination codon to two closely linked AATAAA polyadenylation sites 419 and 434 nucleotides downstream from the TGA terminator. The corresponding 3'-untranslated sequences of the H-2K^d allele are 10 nucleotides shorter. A total of 23 nucleotide differences exist between the two H-2K alleles and, correcting for deletions/insertions (i.e., assuming that a deletion/insertion is a single mutagenic event), 4.2% differences exist between the 3' extragenic regions of the two alleles. In contrast, there are 11.6% differences between the 3'-untranslated regions of the H-2K^b gene and H-2L^d gene over the 310 nucleotide region where homology can be detected. More importantly, beyond this 310 nucleotide region, the H-2L^d and H-2K^d sequences are totally non-homologous (Figure 4). The H-2L^d sequence retains extensive homology with the H-2D^b gene throughout this region. The 3'-flanking DNA of the two H-2K alleles shows extensive homology and exhibits 3% sequence divergence.

Discussion

The sequence of the H-2K^b gene is similar to those of the other H-2 genes reported previously in its general structural features. Comparison of this sequence with the sequence of the H-2K^d allele allows some interesting conclusions to be made.

Comparison of the DNA sequences of homologous genes in different species shows that the introns accumulate mutations at a significantly greater rate than exons (van den Berg *et al.*, 1979; Efstratiadis *et al.*, 1980). For the allele H-2K^b/K^d pair, the converse is true. In fact, the percentage divergence for exons 2 and 3, which encode the polymorphic α_1 and α_2 domains of the H-2 polypeptide is 2–3 times greater than that of the introns.

The sequence divergence of exons 2 and 3 of the two H-2K alleles is the same as the divergence between the non-alleles H-2K^b and H-2L^d; thus, at the DNA level the non-alleles are as similar to each other as are the alleles in this region. In fact, that these two H-2K genes are alleles is apparent only from the sequence homology in the non-coding regions.

Previous analysis of the DNA sequence of allelic and non-allelic globin (Slightom *et al.*, 1980) and immunoglobulin genes (Miyata *et al.*, 1980) has pointed to events resembling gene conversion (this term is deliberately used loosely in this paper) which cause non-alleles to be as similar or more similar to each other than to their respective alleles in part of their DNA sequence.

To evaluate the role of this type of genetic interaction in the generation of polymorphism of H-2 genes we have determined the DNA sequence of the H-2K^b gene and previously compared this with a mutant form of the H-2K^b gene (Flavell *et al.*, 1982b; Weiss *et al.*, 1983). Our previous study of this H-2K^{bml} mutant gene showed that the apparently single mutational event, which caused three amino acid substitutions, has probably resulted from the introduction of a short DNA sequence from another H-2 gene by a mechanism resembling gene conversion. In the comparision of the allelic H-2K^b and H-2K^d genes in this article we have noted 11 possible clustered hot spots of nucleotide substitutions which resemble the hot spot found by us for the H-2K^{bml} mutation both in the length of the DNA region altered and the number of nucleotide substitutions found.

Of these clusters, in three of the cases the novel DNA sequence found in the H-2K allele has a counterpart in another non-allelic H-2 gene. Thus, the codons for residues 81–83 of the H-2K^b gene are identical to those of the H-2D^b gene; codons 155 and 156 of the H-2K^d gene are identical to the H-2L^d gene; and codons 173–177 of the H-2K^b gene are identical to the H-2D^b gene. In all these cases several nucleotides at a given cluster are involved. In another four cases we can detect homologous or identical sequences in other H-2 genes, but since these changes involve, for the most part, fewer nucleotides, their predictive power is less. It should be remembered that there are 20 or more H-2-class I-like DNA segments in the mouse genome (Steinmetz *et al.*, 1982; Flavell *et al.*, 1982) and that only a few of these have been sequenced. Our failure to find a donor sequence for the remaining four cases is therefore not surprising.

It is striking that these clustered differences are mainly restricted to the exons. It is at present not clear what mechanism might be responsible for the introduction of new DNA sequence from donor to acceptor gene. We would point out, however, that this might be mediated by either DNA or RNA and the fact that these changes are seen only in exons so far is consistent with an mRNA mediated genetic exchange.

S. Weaver, M. Edgell and C.A. Hutchinson III (personal communication) have performed an initial sequence comparison of two allelic β-globin genes from the BALB/c and B10 mouse, that is, the same strains compared here. They find a 3–4% sequence divergence for exons 1 and 2 and intron 1 of the two pairs of β-globin alleles. This is the same percentage divergence as the basal level we see for the introns and the non-clustered changes in the exons. However, the majority (60–65%) of the base substitutions between the two H-2K alleles fall within clusters. These clusters are analogous to the changes observed in bml which was isolated as a single mutational event. We believe therefore that micro-gene conversion events are the driving force for the generation of polymorphism in H-2 genes.

Materials and methods

Preparation of cosmid libraries and the cloning of the H-2K^b gene was described previously (Mellor *et al.*, 1982). DNA sequence determination was according to Maxam and Gilbert (1980) and sequence comparisons were performed as described previously (Moschonas *et al.*, 1982).

Acknowledgements

Madlyn Nathanson is gratefully acknowledged for her efforts in preparing this article. E. Weiss was initially supported by a fellowship from the Deutsche Forschungsgemeinschaft. L. Golden was supported by a Helen Hay Whitney Research Fellowship during this work and R. Zakut by an EMBO short term fellowship. Part of this work was supported by the British Medical Research Council and part by Biogen N.V.

References

Efstratiadis,A., Posakony,J.W., Maniatis,T., Lawn,R.M., O'Connell,C., Spritz,R.A., De Riel,J.K., Forget,B., Weissman,S.M., Slightom,J.L., Blechl,A.E., Smithies,O., Baralle,F.E., Shoulders,C.C. and Proudfoot, N.J. (1980) *Cell,* **21**, 653-668.

Evans,G.A., Margulies,D.H., Camerini-Otero,R.D., Ozato,K. and Seidman, J.G. (1982) *Proc. Natl. Acad. Sci. USA,* **79**, 1994-1998.

Flavell,R.A., Bud,H., Bullman,H., Busslinger,M., de Boer,E., De Kleine,A., Golden,L., Groffen,J., Grosveld,F.G., Mellor,A.L., Moschonas,N. and Weiss,E. (1982a) in Bonné-Tamir,B. (ed.), *Human Genetics Part a: The Unfolding Genome,* Alan R. Liss, Inc., NY, in press.

Flavell,R.A., Grosveld,F., Busslinger,M., de Boer,E., Kioussis,D., Mellor, A.L., Golden,L., Weiss,E., Hurst,J., Bud,H., Bullman,H., Simpson,E., James,R., Townsend,A.R.M., Taylor,P.M., Schmidt,W., Ferluga,J., Leben,L., Santamaria,M., Atfield,G. and Festenstein,H. (1982) *Cold Spring Harbor Symp. Quant. Biol.,* in press.

Grosveld,G.C., Rosenthal,A. and Flavell,R.A. (1982) *Nucleic Acids Res.,* **10**, 4951-4971.

Klein,J. (1979) *Science (Wash.),* **203**, 516.

Kvist,S., Roberts,L. and Dobberstein,B. (1983) *EMBO J.,* **2**, in press.

Lalanne,J.L., Bregegere,F., Delarbre,C., Abastado,J.P., Gachelin,G. and Kourilsky,P. (1982) *Nucleic Acids Res.,* **10**, 1039-1040.

MacNight,L.S., Gavis,E.R. and Kingsbury,R. (1981) *Cell,* **25**, 385-398.

Maxam,A.M. and Gilbert,W. (1980) in Grossman,L. and Moldave,K. (eds.), *Methods in Enzymology,* Vol. **65**, Academic Press, NY, pp. 499-560.

Mellor,A.L., Golden,L., Weiss,E., Bullman,H., Hurst,J., Simpson,E., James,R., Townsend,A.R.M., Taylor,P.M., Schmidt,W., Ferluga,J., Leben,L., Santamaria,M., Atfield,G., Festenstein,H. and Flavell,R.A. (1982) *Nature,* **298**, 529-534.

Miyata,T., Vasunaga,T., Yamawaki-Kataoka,Y., Obarta,M. and Honjo,T. (1980) *Proc. Natl. Acad. Sci. USA,* **77**, 2143-2147.

Moore,K.W., Sher,B.T., Sun,Y.H., Eakle,K.A. and Hood,L. (1981) *Science (Wash.),* **215**, 679-682.

Moschonas,N., de Boer,E. and Flavell,R.A. (1982) *Nucleic Acids Res.,* **10**, 2101-2120.

Pease,L.R., Schulze,D.H., Pfaffenbach,G.M. and Nathenson,S.G. (1983) *Proc. Natl. Acad. Sci. USA,* in press.

Reyes,A.A., Schold,M., Itakura,U. and Wallace,R.B. (1982a) *Proc. Natl. Acad. Sci. USA,* **79**, 3270-3274.

Reyes,A.A., Schold,M. and Wallace,R.B. (1982b) *Immunogenetics,* **16**, 1-9.

Slightom,J.L., Blechl,A.E. and Smithies,O. (1980) *Cell,* **21**, 627-638.

Steinmetz,M., Moore,K.W., Frelinger,J.G., Sher,B.T., Shen,F.W., Boyse, E.A. and Hood,L. (1981) *Cell,* **25**, 683-692.

Steinmetz,M., Winoto,A., Minard,K. and Hood,L. (1982) *Cell,* **28**, 489-498.

van den Berg,J., van Ooyen,A., Mantei,N., Schambock,A., Grosveld,G. Flavell,R.A. and Weissmann,C. (1978) *Nature,* **275**, 37-44.

Weiss,E., Mellor, A., Golden,L., Fahrner,K., Simpson,E., Hurst,J. and Flavell,R.A. (1983) *Nature,* in press.