A cDNA clone containing the entire coding sequence of a mouse H-2K\textsuperscript{d} histocompatibility antigen

Jean-Louis Lalanne, Christiane Delarbre, Gabriel Gachelin and Philippe Kourilsky

Unité de Biologie Moléculaire du Gène – ER CNRS no.201 and SC INSERM no.20 – Institut Pasteur, 25 rue du Dr. Roux, 75724 Paris Cédex 15, France

Received 5 January 1983; Revised and Accepted 16 February 1983

ABSTRACT

We have isolated a cDNA clone carrying a 1560 bp long insert which contains the entire coding and 3' untranslated regions of an H-2K\textsuperscript{d} mouse histocompatibility antigen. Its sequence and overall features are described. They point to the existence of unique properties of DNA sequences associated with the H-2K\textsuperscript{d} antigen.

INTRODUCTION

Class I major histocompatibility antigens (H-2 antigens in the mouse) are cell surface glycoproteins involved in a variety of immune phenomena (1, 2). They are heterologous dimers resulting from the non covalent association of a glycosylated 44,000 dalton H-2 heavy chain (3, 4) with a 12,000 dalton polypeptide, β2-microglobulin (2, 5). The H-2 heavy chain spans the cell membrane with its COOH end located within the cytoplasm. In most H-2 haplotypes, serologically-defined H-2 antigens are encoded by genes located in three main loci, named H-2K, H-2D and H-2L (6, 7). Following molecular cloning of the H-2 region, at least 36 H-2 and H-2-related genes have been assigned to the H-2 complex (8, 9), in the H-2\textsuperscript{d} (BALB/c) haplotype. Some of these genes code for serologically defined H-2, Qa and Tla antigens; however most of them code for unidentified gene products related to H-2 molecules which appear to be either anchored to the membrane (9) or secreted into the extracellular environment (10).

Comparatively little is known of the diversity of the mRNAs encoded by members of the H-2 multigene family (11). Only uncomplete H-2 cDNAs have been sequenced (12-15) and only two cDNAs have been identified with transcripts of the H-2K\textsuperscript{b} and H-2D\textsuperscript{b} genes, respectively (16, 17). We have undertaken the cloning of complete H-2 cDNAs in order to identify the processed transcripts (and, thus, the proteins) and to study their expression with the proper vectors. We report here the isolation of a complete H-2 cDNA, identified as a
product of an H-2K<sup>d</sup> gene. The comparison of its sequence with partial protein sequences and other cDNA sequences points to the existence of distinctive traits associated with the K region of the H-2 complex.

**MATERIAL AND METHODS**

**Construction of the cDNA library** - RNAs were extracted from livers of adult DBA/2-Pas mice (haplotype H-2<sup>d</sup>) and fractionated on oligo-dT cellulose as described elsewhere (18). Integrity of the H-2 specific poly A<sup>+</sup> mRNAs was checked by Northern blotting analysis using an H-2 specific cDNA probe (18). 10 µg of poly A<sup>+</sup> mRNA was dissolved in 100 µl of 50 mM Tris HCl buffer, pH 8, 6 mM MgCl<sub>2</sub>, 75 mM KCl, 1 mM DTT, 1 mM dATP, dCTP, dGTP, 10 µCi <sup>32</sup>P-CTP (800 Ci/mM, Amerham) and 40 µg/ml oligo-dT(12-18) (Collaborative Research). The reaction was started by the addition of 100 units of AMV reverse transcriptase (Life Science, St. Petersburg, Florida). The mixture was incubated at 43°C for 45 minutes. Synthesis of the first strand was terminated by addition of 5 µl EDTA, 0.2 M, pH 8. After mRNA was cleaved off by mild alkaline treatment (NaOH 0.3 M, 2 hrs, 50°C), the solution was neutralized and single strand cDNAs were fractionated through a column of Sephadex G150, equilibrated with 5 mM Tris-HCl buffer, pH 7.5, 5 mM NaCl and 0.1 % SDS. Material eluted with the void volume was pooled and precipitated with ethanol in the presence of 0.3 M sodium acetate buffer, pH 5.6. cDNAs were thenafter tailed with dCTP using deoxynucleotidylic terminal transferase (19). Synthesis of the second strand was primed with oligo-dG(12-18) (Collaborative Research) and continued as described elsewhere (19). Tailing of double stranded cDNAs and PstI-cleaved pBR237, plasmid-cDNA hybridization and E. coli transformation, were carried out according to Auffray et al (20).

**Screening of the cDNAs** - Colonies harboring recombinant cDNA were screened by in situ hybridization (21) using the 600 bp long HindII fragment of an H-2 cDNA described previously (pH-2<sup>d</sup>-4) (14). 40 clones were found positive out of a total of 25000 colonies.

Plasmid DNA was prepared from 1 ml cultures and screened for their size. Several clones harboring inserts longer than 1000 bp were retained for further analysis. Large batches of plasmid DNA were prepared according to conventional procedures (22, 23).

After construction of the restriction maps (Fig. 1), clone pH-2<sup>d</sup>-33 was found to possess the SacII restriction sites abundant in the 5' part of H-2<sup>d</sup> genes (24, 25) and was retained as a good candidate for being a full length
Figure 1: Restriction map of plasmid pH-2d-33. The 1560 bp long insert (including poly A) was drawn as a straight line thicker in the coding region (C) than in the non-coding region (NC). It is flanked on both sites by GC tails of about 30 bp (shown as a thinner line) then, pBR327 sequences. The restriction map was constructed as described (14). Restriction sites indicated in the map are those used for sequencing. The sequencing strategy is indicated by arrows. Due to the lack of convenient restriction sites, a small part of the non-coding region was sequenced several times off the same strand. The remainder of the sequence was determined from both strands. All labelings were carried out on the 5' ends by γ exchange with the exception of a labeling at the 3' end using cordycepin (*).

H-2 cDNA. Its DNA sequence was determined using the Maxam and Gilbert technique (26) following the strategy shown in Fig. 1. The recombinant plasmid was found to bear a deletion between nucleotides 2524 and 3216 in the pBR327 sequence, therefore lacking the reconstructed PstI site to the left of the insertion.

Materials - The plasmid used in the present studies was pBR327, a deletion derivative of pBR322 (27). All other reagents and materials have been specified in previous publications from this laboratory (11-14).

RESULTS

General features of clone pH-2d-33 - The complete sequence of clone pH-2d-33 is shown in figure 2. The first ATG opening the longest open reading frame is preceded at its 5' end by a stretch of 25 nucleotides (excluding the GC tail). The translation of the 1104 nucleotides of the open reading frame is given in figure 3, along with the complete protein sequence of H-2Kd and that of fragments of the amino acid sequences of H-2Kd, Dd and Ld (review in 4 and 28). The coding region is followed at its 3' end by a 425 bp-long non-coding
Figure 2: Complete nucleotide sequence of clone pH-2^d^-33.
Figure 3: The deduced amino acid sequence of clone pH-2^d^-33 is aligned with the sequences of H-2K^d^- (4), H-2K^b^- (3) and the fragments of sequence of H-2D^d^- and H-2L^d^- (4). The arrows indicate the boundaries between exons as deduced from the sequences of H-2L^d^- (25, 29) and H-2K^d^- (33) genes. Putative glycosylation and phosphorylation sites are indicated by (●) and (O) respectively. The amino acids considered to be characteristic of the products of the K region are underlined.
sequence. The total length of the cDNA is 1560 bp. Assuming a poly A tail of 100 to 200 residues at its 3' end, this size is compatible with that of intact H-2 mRNAs as determined by the Northern technique (18).

**General organization of the protein encoded by pH-2^d^-33** - The NH₂ terminal amino acid sequences of the few H-2 antigens sequenced can be aligned with the amino acid sequence deduced from the nucleotide sequence of the cDNA, starting from amino acid number 22 (glycine). The first 21 amino acids coded by pH-2^d^-33 should therefore correspond to the signal sequence of the H-2 precursor, removed during in vivo synthesis and processing (2, 29). This stretch is composed of hydrophobic amino acids, with the exception of the 20th residue (Arginine). The size of this putative signal sequence is compatible with that of the fragment removed during in vitro processing of pre-H-2 proteins (29).

The best alignment of the 63 nucleotides coding for the leader sequence with the corresponding region of the H-2L^d^ gene(s) (25, 30) is obtained through three insertions/deletions (which do not, however, alter the hydrophobic character of the leader sequences). Only seven differences in nucleotide sequence were identified between these two stretches of 63 nucleotides thus pointing to a striking degree of homology between these two sequences, corresponding to a 90% homology.

Since glycine is the first amino acid of the K, D, D, K sequences, the signal sequence of the protein coded by pH-2^d^-33 is likely to be removed by proteolytic cleavage of the precursor between Ala 21 and Gly 22. Downstream from the region coding for the signal sequence we identified a region coding for a molecule homologous to all H-2 antigens which have been sequenced to date (Fig. 3). The best fit of the sequence encoded by pH-2^d^-33 corresponds to H-2K^d^; over the first 90 amino acids, only three differences are noted with K^d^, contrasting with 23 and 24 with D^d^ and L^d^ antigens respectively.

With one exception, all amino acid sequences common to all H-2K gene products (4) were identified in pH-2^d^-33. These common sequences are AIVT (296-299) and TSDL (329-332) (using the one letter code for amino acids). The LITKH sequence (141-145) is missing and replaced by LITRR. Methionine residues are found at positions 52, 98, 228, 307, 309; the size of the fragments obtained from H-2K^d^ molecules upon cyanogen bromide treatment assigns methionine residues at positions 52, 98, 228, 309 (as in the protein coded by pH-2^d^-33) some of which missing from L^d^ and D^d^ gene products (31). Cystein residues are found at positions 102-104, 203-259, thus allowing the formation of disulphide bridges within the second, and the third domain of the protein respectively. Such
bridges exist in all H-2 molecules (4). Glycosylation sites are found at positions 86 and 176; in fact, H-2Kd is known to be glycosylated at these positions (4). A third glycosylation site exists at position 256, as it does in the protein encoded by clone pH-2d-4 (14). However, its location in the vicinity of the disulphide bridge of the third domain might prevent glycosylation at this site, since it appears to be buried within the protein globule.

The transmembrane hydrophobic region extends from valine 285 to methionine 309. It is ended, on its cytoplasmic site, as already noted for other H-2 clones (12) by a short stretch of positively-charged amino acids, probably involved in the correct positioning of the chain within the membrane. A distinctive feature of H-2Kd antigens is the presence of only two arginine residues at the end of the transmembrane region, instead of three in all other known H-2 molecules (4). It should be stressed that pH-2d-33 possesses two arginines at this location.

The cytoplasmic region of the H-2Kd molecule is 38 amino acid long. It is ended by the stretch of 9 amino acids (VMVHDPHLA) so far characteristic of products of the H-2K locus (4, 28).

**Organization of the non-coding regions** - The open reading frame of pH-2d-33 is preceded by 25 nucleotides corresponding to a part of all of the 5' non-coding region. This sequence has no obvious homology with that of the corresponding region of the H-2Ld gene(s) (25, 30) but is identical to the corresponding region of an H-2Kd gene (see below).

The 3' non-coding region of the cDNA comprises 426 nucleotides (excluding the poly A tail). This non-coding region is 99.8% homologous to that of clone pH-2d-4 (14) (one difference out of 426 nucleotides). Clone pH-2d-33 lacks the 3' non-coding, highly repeated sequence found in most other cDNAs (14, 15). However, the 3' end of pH-2d-33 is shorter by 20 nucleotides as compared to that of the similar clone pH-2d-4 (14). This is probably due to the fact that the polyadenylation signal used in clone pH-2d-33 is the first of the two found in pH-2d-4 (14) (Fig. 4).

**DISCUSSION**

The identification of clone pH-2d-33 as encoding an H-2Kd histocompatibility antigen rests on several lines of evidence: first, the deduced amino acid sequence fits the fragmentary data on H-2Kd primary structure. Second, the 3' untranslated region of pH-2d-33 contains the NC2 sequence of pH-2d-4 (14) shown to be specific of the K region (32). Third, a Kd gene of DBA/2 mice has recently been isolated and its sequence was kindly made available to
Figure 4: Clones pH-2^d^-4 (14) and pH-2^d^-33 are identical over a large stretch of their 3' non-coding regions but differ at their 3' end. The first of the two polyadenylation sites (underlined) found in pH-2^d^-4 seems to be used in pH-2^d^-33, thus yielding a fully processed poly A^+ H-2Kd mRNA lacking the last 18 nucleotides.

This identification will be considered definitive only after pH-2^d^-33 has been expressed and its products detected by proper antisera. There exist three amino acid differences among the available amino acid sequences of H-2Kd and that of the protein encoded by pH-2^d^-33, at positions 43 (alanine → proline),
67 (valine + alanine) and 78 (threonine + leucine). It is not known whether these changes can alter the overall structure of the protein. They may simply reflect evolutionary drifts within inbred populations of mice: pH-2\textsuperscript{d-33} derives from mRNAs extracted from the liver of DBA/2 mice whereas the protein sequences have been determined mostly with material prepared from BALB/c mice. Although the two strains have retained the same H-2\textsuperscript{d} haplotype, they have bred separately for about 50 years. Small differences which would not result in immunological alterations can, therefore, have accumulated within the same H-2K\textsuperscript{d} gene.

We have previously reported the sequence of pH-2\textsuperscript{d-4}, an uncomplete H-2 cDNA. Its sequence and features were compatible with its identification as an H-2K\textsuperscript{d} cDNA (14). The comparison of the sequence of this clone with pH-2\textsuperscript{d-33} shows 12 nucleotide changes: seven of them result in amino acid changes, four are silent and one is in the 3' non coding region. Strikingly, most of the differences have accumulated over the first half of the third domain (Fig. 5), whereas membrane and cytoplasmic regions remained nearly identical. Again, it should be noted that pH-2\textsuperscript{d-4} derived from SL2 lymphoma cells, of DBA/2 background. SL2 cells have been maintained in continuous passage for many years. Their H-2 genes may have evolved faster than the corresponding genes in the animals themselves, due to the large number of SL2 cell divisions, relative to the small number of generations of DBA/2 mice during the same time period. If this hypothesis were to be correct, then hot spots of genetic changes, such as in the third domain, are likely to exist in H-2 genes.

Southern blot analysis of DNA from various recombinant strains of mice (32) showed that genes possessing the 3' non coding region of pH-2\textsuperscript{d-5} (and thus also of pH-2\textsuperscript{d-4} and -33) derived from the K end of the H-2 complex in the d and b haplotypes. It cannot be ascertained from these data, however, if there exists only one single H-2K\textsuperscript{d} gene. Various homologous genes could exist which yield the same restriction fragments. In fact, recent serological data suggest the existence of different products from the K\textsuperscript{d} region in the H-2\textsuperscript{d} haplotype (35). However, the description of overlapping cosmid clones in the H-2K\textsuperscript{d} region makes the hypothesis of a multiplicity of H-2K\textsuperscript{d} genes unlikely (8, 9).

The results presented here tend to suggest that there exists, at least in the b and d haplotypes, several traits distinctive of the K region and not found in genes located in other regions (D, L, Qa, etc...). These include the presence of a distinctive 3' non coding region lacking the repeated sequence; the presence of distinctive short amino acid stretches within the intracellu-
lar regions; the presence of an additional tail at the COOH part of the molecule (4, 14, 17). The latter feature of K gene products may be related to functions of H-2K antigens particularly with respect to their association with the cytoplasmic region and with cytoskeleton. The possibility now offered to express complete H-2 cDNAs before and after having made alterations in the region of the sequence coding for the cytoplasmic part of the H-2K molecule will allow to study this hypothesis.

ACKNOWLEDGEMENTS

We thank Dr. M. Cochet for many helpful discussions, Drs. P. Baldacci and D. Morello for help and F. Plata for his critical reading of the manuscript.

This work was supported by grants from C.N.R.S. (E.R. 201 ; A.T.P. 4246 and 5039), I.N.S.E.R.M. (S.C. 20 and P.R.C. 124031) and the Fondation pour la Recherche Médicale Française.

J.-L.Lalanne was supported by a fellowship from Elf-Aquitaine Company

REFERENCES