Proc. Natl. Acad. Sci. USA
Vol. 79, pp. 893–897, February 1982
Immunology

Exon/intron organization and complete nucleotide sequence of an *HLA* gene

(genomic clone/variability/intracellular segment)

MARIE MALISSEN, BERNARD MALISSEN, AND BERTRAND R. JORDAN

Centre d'Immunologie, Institut National de la Santé et de la Recherche Médicale, Centre National de la Recherche Scientifique de Marseille-Luminy, Case 906-13288, Marseille Cedex 9-France.

Communicated by Baruj Benacerraf, October 19, 1981

ABSTRACT **We have isolated and determined the sequence of a genomic clone containing the gene for a human class I transplantation antigen. The gene contains seven exons. The first five exons code respectively for a signal peptide, for the first, second, and third extracellular domains of the protein molecule, and for the transmembrane region. The cytoplasmic segment is encoded by part of the fifth and the sixth and the seventh exons. The structure of the protein encoded by this unit is closely homologous with known class I transplantation antigens.**

Knowledge of the structure of the polymorphic class I transplantation antigens encoded in the major histocompatibility complex (*MHC*; *HLA-A*, *-B*, and *-C* in man, *H-2K*, *D*, and *L* in mouse) has made rapid progress recently. Complete protein sequences have been obtained by radiochemical sequence determination for mouse antigens (1); in the human MHC, the structure of the extracellular portion and cytoplasmic segment of two molecules, HLA-A2 and HLA-B7, has been determined (2, 3); partial data are also available on the transmembrane portion of the molecule. These results strengthen the assumption of a domain-like organization (4, 5) of class I antigens and have suggested possible variable regions in these molecules. More recently, advances have been made toward isolation of the corresponding genes; cDNA clones have been obtained in both the *H-2* (6, 7) and the *HLA* (8, 9) systems and used to probe the organization of these genes by Southern blotting (7, 10) and to isolate genomic clones (11, 12).

We have recently isolated HLA genomic clones from a human DNA library in λ charon 4A, by using the mouse *H-2* cDNA probes obtained by Kvist *et al.* (6). One of these clones, called λ HLA 12, was shown to contain sequences coding for the third domain of the HLA molecule (11). We have now studied this clone in more detail; a 5.6-kilobase (kb) *Hin*dIII fragment was found to contain one complete *HLA* gene, and the sequence of this gene has been completely determined. These results reveal the general organization of an *HLA* transcription unit, support the domain model for class I antigens, provide some evidence against the involvement of DNA rearrangements in the expression of these molecules, and show the unexpected complexity of the cytoplasmic coding region. They also provide a new complete class I antigen structure that can be compared with previously determined protein sequences.

MATERIALS AND METHODS

Enzymes and Other Reagents. Restriction enzymes were obtained from Bethesda Research Laboratories, Gaithersburg, MD, New England Biolabs, and Boehringer Mannheim. Terminal transferase, polynucleotide kinase, and bacterial monophosphoesterase were from Bethesda Research Laboratories or P-L Biochemicals. Radiochemicals were obtained from Amersham.

DNA Labeling and Fragment Preparation. 3′- and 5′-end labeling (with or without prior denaturation as appropriate) was carried out on either restriction digests or fragments eluted from acrylamide or agarose gels. Uniquely labeled fragments were then obtained by secondary restriction enzyme digestion, strand separation on 5% or 10% neutral acrylamide gels or both.

DNA Sequence Analysis. Sequence determination was carried out according to Maxam and Gilbert (13); reaction products were analyzed on 20%, 8%, and 6% thin acrylamide/urea gels. Essentially all the sequence was determined independently on both strands, using either different restriction sites or 3′- and 5′-end labeling at the same site; overlapping sequences were obtained for all restriction sites except the *Taq* I site present ahead of the first exon at position 100 and the *Sst* I site in the transmembrane–first cytoplasmic region exon.

RESULTS

Isolation of Genomic Clones. A screen by plaque hybridization of the human DNA library in phage λ charon 4A constructed by Lawn *et al.* (14) was carried out with a mouse *H-2* cDNA probe, the H2.1 fragment from the pH-2^d-3 plasmid (11, 15), containing the transmembrane, cytoplasmic, and 3′ noncoding sequences of the corresponding mRNA. From 10^5 phage plaques, corresponding to approximately half a human haploid genome, 15 confirmed positive phages were obtained. Most of these clones are different from each other (unpublished). This suggests a surprisingly high number of coding sequences in human DNA, perhaps 30 per haploid genome, and is not inconsistent with results recently obtained in the *H-2* system by Southern analysis of genomic DNA (7, 10). The λ HLA 12 clone was chosen for further study, mapped, and shown by sequence analysis to contain authentic *HLA* coding sequences (11).

Sequence Determination Strategy. Subclone pHLA 12.4—the 5.6-kb *Hin*dIII fragment from clone λ HLA 12, which hybridized with both third domain and transmembrane-cytoplasmic mouse probes (15) and was recloned in plasmid pBR328 (16)—is shown in Fig. 1. A complete set of *HLA* coding sequences was found in this segment. Its organization was determined mainly by DNA sequence determination and comparison of the possible translation products with the HLA-B7 and -A2 protein sequences.

This strategy allowed us to determine the organization of this transcription unit and to show that it contains all the expected coding sequences; 4.1 kb of DNA, containing the complete set

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

Abbreviations: MHC, major histocompatibility complex; kb, kilobase(s).

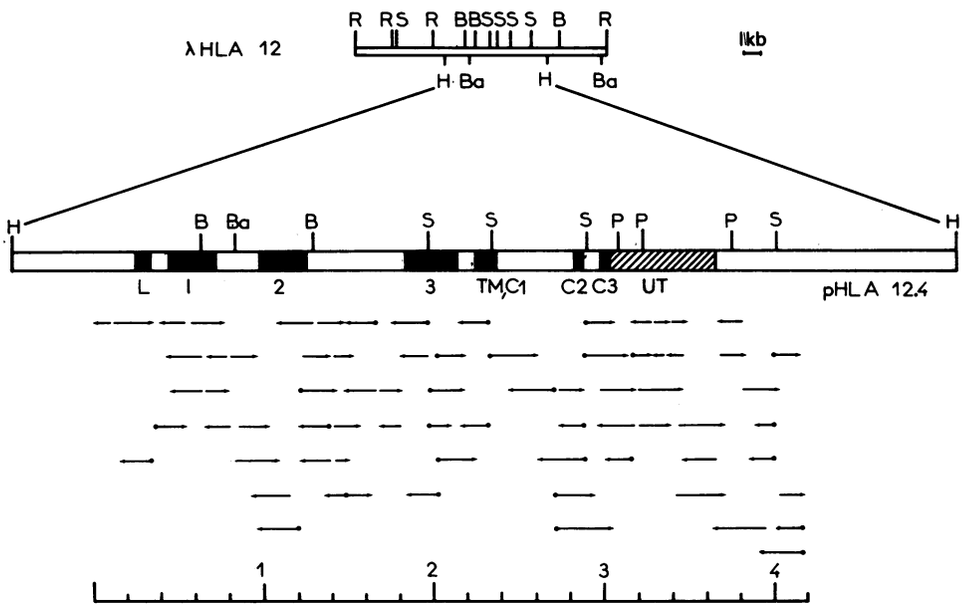

FIG. 1. Maps of genomic clone λ HLA 12 and the pHLA 12.4 subclone. (*Top*) Partial restriction map of the 15-kb insert present in clone 12. The *Eco*RI sites at the end of the insert were generated during construction of the library (14). The 5.6-kb *Hin*dIII fragment was subcloned in plasmid pBR328 (20) and studied by Southern blotting and DNA sequence analysis. B, *Bgl* II; Ba, *Bam*HI; R, *Eco*RI; S, *Sst* I; P, *Pvu* II. (*Bottom*) Map of subclone pHLA 12.4 (5.6-kb *Hin*dIII fragment from clone 12), ■, Positions of exons; ▨, 3′ Untranslated region. Arrows indicate individual sequences obtained by the Maxam–Gilbert procedure; those with a dot indicate 3′ labeling of the DNA strand, others indicate 5′ labeling.

of exons, was analyzed by the Maxam–Gilbert chemical degradation procedure (13). The sequence determination strategy is summarized in Fig. 1 and the DNA sequence obtained is shown in Fig. 2. The locations of the exons and introns are shown in Fig. 1 and correlated with the HLA protein in Fig. 3. The location of the exon/intron boundaries could be determined from comparison of the possible translation products of the DNA sequences with the HLA-B7 protein sequences; G-T (beginning of intron) and A-G (end of intron) dinucleotides (17) were found in every case at the point at which the protein sequence deduced from the DNA became unrelated to the HLA-B7 sequence. A check of these locations was provided by the similarly determined location of the starting point for next exon. One example is shown for the first-domain/intron/second-domain junction in Fig. 4.

A Complete *HLA* Gene is Found in Genomic DNA. It has been suggested that class I antigens could be synthesized from a small set of third-domain coding regions and a large set of "variable" first- and second-domain regions, assembled through DNA rearrangements or differential RNA splicing (18). The finding of a complete *HLA* transcription unit containing coding sequences for all domains argues against the second version of this hypothesis, while the invariability of DNA blots from sperm or liver cells when probed with *H-2* cDNA probes (7) provides evidence against the first version.

A Probable Signal Peptide Exon is Found 125 Base Pairs Upstream of the First Domain Exon. A signal sequence containing ≈20 amino acids is known to exist in the initial translation product of *HLA* mRNA (19). The corresponding region could not be identified directly because the protein sequence has not been reported. However, we found a putative signal peptide exon (see Figs. 2 and 5) that starts with a methionine codon, encodes a stretch of 21 amino acids, 18 of which are hydrophobic or uncharged, and contains, at its 3′ end, a "beginning of intron" G-T dinucleotide that correlates with an "end of intron" A-G dinucleotide in the first codon of exon 2 (first domain). No other such structure is found in the 433 nucleotides whose sequences have been determined upstream of the first domain exon.

The First, Second, and Third Domains and the Transmembrane Segment of the HLA Antigen are Coded by Distinct Exons. The exon/intron organization of the pHLA 12.4 transcription unit in this region follows closely (and confirms) the domain structure proposed for class I transplantation antigens (4, 5): the first, second, and third domains and the transmembrane segment are encoded by distinct exons (exons 2, 3, 4, and 5, respectively) separated by introns of various lengths. This type of organization is similar to that found for immunoglobulins and suggests that these domains may have been assembled after evolving independently, as originally suggested by Gilbert (20). Exon 5 deserves special attention because it actually codes for first a short hydrophilic segment (Glu-Pro-Ser-Ser), which could be considered as part of the third domain, then a stretch of 28 hydrophobic amino acids, which constitute the transmembrane segment, and finally a hydrophilic (Met-Trp-Arg-Lys-Lys-Ser-Ser) peptide, which is the beginning of the cytoplasmic region and is considered to play a role in anchoring the antigen in the cell membrane (for review of the data on this region, see ref. 1).

The Cytoplasmic Segment is Split Between Three Exons. We were surprised to find that the cytoplasmic segment coding sequences are found in three distinct exons: the first one, 7 amino acids long, is continuous with the transmembrane exon; the second one, 438 nucleotides further to the 3′ end of the gene, is 11 amino acids long and separated by a 141-nucleotide-long intron from the last, 14-amino acid-long exon, which is contiguous with the 3′ untranslated region. Recent results in the mouse system indicate a similarly complex gene structure in this region (12) together with differences in the cDNA sequence that suggest the existence of alternative processing pathways (12, 15). Although such data are not yet available in the human system, the similarity in genetic organization of the cytoplasmic coding regions indicates that alternative splicing should also be considered in this case. It is clear that much remains to be understood on the structure and function of the intracellular portion of transplantation antigens.

Comparison of the Protein Sequence Encoded by pHLA 12.4 Confirms Regions of Variability Detected by HLA-B7 Versus HLA-A2 Comparisons and Suggests Additional Variable Regions. The sequence of the protein encoded by the pHLA 12.4 exons is compared with the HLA-B7 protein in Fig. 5. In the first domain, a small cluster of differences is seen between positions 74 and 85. This cluster was apparent in HLA-B7 versus HLA-A2 comparisons. Altogether, the first domain of pHLA 12.4 shows 16 differences with the B7 protein and 22 with the A2 protein. In the second domain, 15 differences are

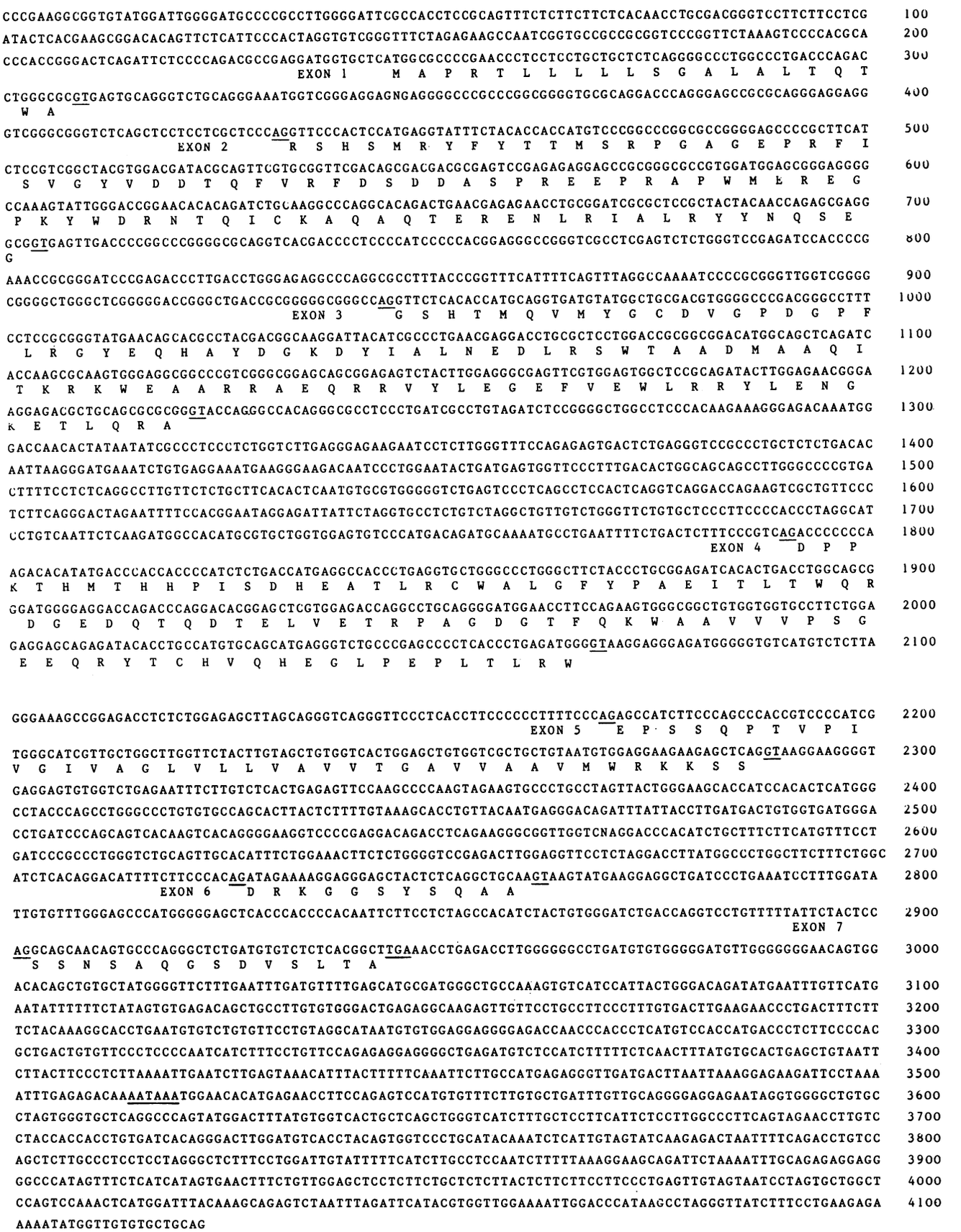

FIG. 2. DNA sequence of a 4.1-kb segment of pHLA 12.4 containing a complete *HLA* gene. Amino acid sequences encoded by the exons of pHLA 12.4 are shown below the DNA sequence. Splicing signals, the termination codon, and the polyadenylylation site are underlined. Amino acids: A, Alanine; C, Cysteine; D, aspartic acid; E, glutamic acid; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine; S, serine; T, threonine; V, valine; W, tryptophan; Y, tyrosine.

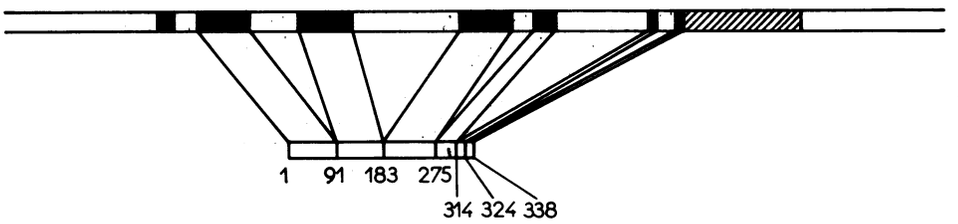

FIG. 3. Organization of exons and introns in pHLA 12.4. The genomic structure is correlated with the protein structure of a class I transplantation antigen. All exons code for sequences highly homologous to both the HLA-B7 and the -A2 molecules (see Fig. 5).

found with B7, 18 with A2, with some clustering around residues 110 (not seen in A2 versus B7 comparisons), and 175 (corresponding to a detected A2/B7 variability region). One of these differences may have important implications: the cysteine normally found at position 164, which is involved in the second-domain disulfide bridge, is not found in the sequence. A phenylalanine is found instead; this corresponds to a one-base change (TGC to TTC). Thus, the protein encoded by the gene whose sequence we have determined cannot form the disulfide bridge found in the second domain of all transplantation antigens whose sequences have thus far been determined. The third domain is, as expected, similar to the A2 and B7 sequences with a few differences (i.e., four with B7 and eight with A2). The transmembrane segment displays a number of differences (the B7 sequence is not completely established in this region), although most of these are Ala-Val-Leu replacements (i.e., the hydrophobic character of the protein sequence is conserved). Finally, the cytoplasmic segment is similar to the A2 and B7 sequences (six differences with B7 and one with A2).

Altogether, those variability regions tentatively defined on the basis of A2/B7 (4) comparisons are also evident in pHLA 12.4/A2 or B7 comparisons; another possible cluster of differences is also found in the second exon. Thus, the delineation of possible variable regions in HLA molecules (apart from the general feature that the third exon is more conserved than the others) will require determination and comparison of a large sample of sequences.

DISCUSSION

Comparison With *H-2* Gene Structure. Mouse *H-2* genomic clones have recently been isolated by Steinmetz *et al.* (12). A complete sequence was derived for an *H-2* gene contained in one of these clones. This is considered to be a pseudogene because it contains two termination codons in transmembrane and cytoplasmic exons and also a charged residue, aspartic acid, inside the transmembrane domain. It can reasonably be assumed, however, that its exon/intron organization is representative of functional *H-2* transcription units. The arrangement reported by Steinmetz *et al.* is in fact strikingly similar to our results except at the end of the cytoplasmic region where the mouse gene contains an additional intron just before the 3′ untranslated sequence.

Is pHLA 12.4 a Gene or a Pseudogene? Pseudogenes have been found in the 5S gene family (21), among globin genes (22), and recently in the mouse *H-2* cluster (12). The possibility that the pHLA 12.4 segment is in fact a pseudogene must be discussed. As mentioned above, the second cysteine involved in the disulfide bridge present in the second domain is replaced by a phenylalanine residue. This raises doubts about the structural rigidity of the second domain of the protein encoded by this gene and its ability to express serological allotypes and the usual functions of MHC antigens.

We have not detected transcription initiation signals upstream of the signal peptide coding sequence. Given the size of *HLA* RNA [1650 nucleotides (18)] and of the 3′ noncoding region (nearly 600 nucleotides), this would indicate the existence of an intron ahead of this coding sequence, as in the case of the ovalbumin gene (23). pHLA 12.4 seems correctly arranged because we have not encountered any termination codon in the exons, all the exon/intron junctions are provided with the requisite splicing signals (16), and the protein corresponding to this gene is closely homologous to known transplantation antigens (3, 4).

Which Part of the *MHC* Gene Cluster Does pHLA 12.4 Come From? The human DNA library from which this clone was isolated was constructed with placental DNA (14) and the corresponding *HLA* typing is unknown.

Although the DNA sequence we have determined provides the complete structure of the protein that could be coded by this gene, this does not allow us to infer the corresponding genetic locus. On the basis of available sequence data, there is no evidence for "*A*-ness" or "*B*-ness" (24); i.e., there is no feature of the structures that assigns a particular sequence to a particular allelic series of class I loci. In fact, the putative pHLA 12.4 protein is closer to the B7 product than to the A2 molecule in its extracellular domains while the reverse is true for the cytoplasmic region. The mouse cDNA probe used for isolation of the clone, which covers transmembrane, cytoplasmic, and 3′ noncoding sequences, could have picked up any class I gene or even other genes as long as they bear some resemblance in the relevant region. The gene structure which we have determined does show that the DNA segment studied codes for a molecule of the class I type; however, we do not know whether this originates from the *HLA-A*, *-B*, or *-C* regions (25) or from the possible human counterpart of the mouse *TL-Qa* region (26, 27, 28).

In conclusion, this study provides the basic structure of a human class I antigen gene. In addition to confirming the domain model for the extracellular part of the protein and strengthening the case for nonrearrangement of these genes, it indicates peculiar features for the cytoplasmic region that suggest the possibility of alternative splicing pathways. In the re-

```
87          90                          87          91
Q  S  E  A  G  S  H  T      Q  S  E  A  G  S  H  T      HLA B7 protein sequence
Q  S  E  G (G)(E)(L)(T)    (G)(G)(G)(Q) G  S  H  T      pHLA 12.4 translation
CAG AGC GAG GGC GGT GAG TTG ACC.....GGG GGC GGG CAA GGT TCT CAC ACC   pHLA 12.4 DNA sequence
Exon 2          ←    Intron (245 bp)          →    Exon 3
```

FIG. 4. Determination of exon/intron junctions by comparison with the HLA-B7 protein sequence. The pHLA 12.4 DNA sequences at the end of the second exon (first domain) and at the beginning of the third exon (second domain) are shown, and one of the possible translation products is compared with the HLA-B7 protein sequence in this region (i.e., amino acids 87–94). The homology between the two sequences breaks down and reappears at locations that correlate with the presence of G-T ("beginning of intron") and A-G ("end of intron") dinucleotides, thus locating the exon/intron boundaries. The single-letter amino acid code is given in Fig. 2.

```
Signal peptide exon

1         10         20
MAPRTLLLLLSGALALTQTWA

First domain exon

1         10         20         30         40         50         60         70         80         90
RSHSMRYFYTTMSRPGAGEPRFISVGYVDDTQFVRFDSDDASPREEPRAPWMEREGPKYWDRNTQICKAQAQTERENLRIALRYYNQSEG
G          SV    R                          A           I Q   E         Y       D  S  NLRG       A

Second domain exon

         100        110        120        130        140        150        160        170        180
GSHTMQVMYGCDVGPDGPFLRGYEQHAYDGKDYIALNEDLRSWTAADMAAQITKRKWEAARRAEQRRVYLEGEFVEWLRRYLENGKETLQRA
    L S           RL   HD Y                          T     Q        E     A     C              DK E

Third domain exon

       190        200        210        220        230        240        250        260        270
DPPKTHMTHHPISDHEATLRCWALGFYPAEITLTWQRDGEDQTQDTELVETRPAGDGTFQKWAAVVVPSGEEQRYTCHVQHEGLPEPLTLRW
       V                                                    R  E                              K

Transmembrane and cytoplasmic exons

    280        290        300        310            320            330      338
EPSSQPTVPIVGIVAGLVLLVAVVTGAVVAAVMWRKKSS      DRKGGSYSQAA      SSNSAQGSDVSLTA
     S            AV AV           C R       GG              C D
```

FIG. 5. Comparison of the protein encoded by the *pHLA 12.4* gene and the HLA-B7 transplantation antigen. The protein sequence deduced from our data is shown. Differences with the HLA-B7 protein are indicated below. The single-letter amino acid code is given in Fig. 2.

gion coding for the extracellular part of the molecule, the structure of the gene closely parallels the functional organization of the transplantation antigen. It seems likely that the same relationship holds for the intracellular region: indeed, if the genetic complexity of the cytoplasmic region indicates a corresponding functional complexity for the intracellular segment of the protein, we are left to wonder whether the part of the transplantation antigen located *inside* the cell may not be extremely important.

Note Added in Proof. *In vitro* transcription experiments have shown recently that a major transcript is initiated from pHLA 12.4 DNA upstream of the signal sequence (C. Kedinger and P. Chambon, personal communication).

We thank Dr. F. M. Kourilsky for enthusiastic support and many discussions; Drs. P. Kourilsky, F. Bregegere, and B. Cami for advice and information; Dr. R. Staden (Medical Research Council, Cambridge) for computer programs; Prof. J. Kern (Centre de Recherches sur les Mecanismes de la Croissance Cristalline, Marseille) for computing facilities; and Drs. L. Hood and S. Weissman for communication of their results prior to publication. This work was supported by the Institut National de la Santé et de la Recherche Médicale (CRL 80 1 027) and the Delegation Generale a la Recherche Scientifique (ACC 80 E 0871). M.M. was the recipient of a Delegation Generale a la Recherche Scientifique training grant.

1. Nathenson, S. G., Uehara, H., Ewenstein, B. M., Kindt, T. J. & Coligan, J. E. (1981) *Annu. Rev. Biochem.* **50**, 1025–1051.
2. Orr, H. T., Lancet, D., Robb, R. J., Lopez de Castro, J. A. & Strominger, J. L. (1979) *Nature (London)* **232**, 266–270.
3. Orr, H. T., Lopez de Castro, J. A., Parham, P., Pleogh, H. L. & Strominger, J. L. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 4395–4399.
4. Lopez de Castro, J. A., Orr, H. T., Kostyk, T., Mann, K. L. & Strominger, J. L. (1979) *Biochemistry* **18**, 5704–5711.
5. Coligan, J. E., Kindt, T. J., Uehara, H., Martinko, J. & Nathenson, S. G. (1981) *Nature (London)* **291**, 35–39.
6. Kvist, S., Bregegere, F., Rask, L., Cami, B., Garoff, H., Daniel, F., Wiman, K., Larhammar, D., Abastado, J. P., Gachelin, G., Peterson, P. A., Dobberstein, B. & Kourilsky, P. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 2772–2776.
7. Steinmetz, M., Frelinger, J. G., Fisher, D., Hunkapiller, T., Pereira, D., Weissman, F. M., Uehara, H., Nathenson, S. G. & Hood, L. (1981) *Cell* **24**, 125–134.
8. Ploegh, H. L., Orr, H. T. & Strominger, J. L. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 6081–6085.
9. Sood, A. K., Pereira, D. & Weissman, S. M. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 616–620.
10. Cami, B., Bregegere, F., Abastado, J. P. & Kourilsky, P. (1981) *Nature (London)* **291**, 673–675.
11. Jordan, B. R., Bregegere, F., & Kourilsky, P. (1981) *Nature (London)* **290**, 521–523.
12. Steinmetz, M., Moore, K. W., Frelinger, J. G., Taylor Sher, B., Shen, F. W., Boyse, E. A. & Hood, L. (1981) *Cell* **25**, 683–692.
13. Maxam, A. M. & Gilbert, W. (1980) *Methods in Enzymology* **65**, 499–560.
14. Lawn, R. M., Fritsch, E. F., Parker, R. C., Blake, G. & Maniatis, T. (1978) *Cell* **15**, 1157–1174.
15. Bregegere, F., Abastado, J. P., Kivst, S., Rask, L., Lalanne, J. L., Garoff, H., Cami, B., Wiman, K., Larhammar, D., Peterson, P. A., Gachelin, G., Kourilsky, P. & Dobberstein, B. (1981) *Nature (London)* **292**, 78–81.
16. Soberon, X., Covarrubias, L. & Bolivar, F. (1980) *Genetics* **9**, 287–305.
17. Lewin, B. (1980) *Cell* **22**, 324–326.
18. Bodmer, F. W. (1981) *Tissue Antigens* **17**, 9–20.
19. Ploegh, H. L., Cannon, L. E. & Strominger, J. L. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 2273–2277.
20. Gilbert, W. (1978) *Nature (London)* **271**, 501.
21. Miller, J. R., & Brownlee, G. G. (1978) *Nature (London)* **275**, 556–558.
22. Proudfoot, N. J. & Maniatis, T. (1980) *Cell* **21**, 537–544.
23. Gannon, F., O'Hare, K., Perrin, F., Le Pennec, J. P., Benoist, C., Cochet, M., Breathnach, R., Royal, A., Cami, B. & Chambon, P. (1979) *Nature (London)* **278**, 428–434.
24. Ploegh, H. L., Orr, H. T. & Strominger, J. L. (1981) *Cell* **24**, 287–299.
25. Bach, F. H. & Van Rood, J. J. (1976) *N. Engl. J. Med.* **295**, 806–813.
26. Gazit, E., Terhorst, C. & Yunis, E. J. (1980) *Nature (London)* **284**, 275–277.
27. Gazit, E., Terhorst, C., Mahoney, R. J. & Yunis, E. J. (1980) *Hum. Immunol.* **1**, 97–109.
28. Cotner, T., Mashimo, H., Kung, P. C., Golstein, G. & Strominger, J. L. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 3858–3862.