Proc. Natl. Acad. Sci. USA
Vol. 83, pp. 1428–1432, March 1986
Genetics

Complete sequence of HLA-B27 cDNA identified through the characterization of structural markers unique to the *HLA-A*, *-B*, and *-C* allelic series

(major histocompatibility complex/disease associations/DNA sequence homology/gene duplication)

HANNELORE SZÖTS*, GERT RIETHMÜLLER*, ELISABETH WEISS*, AND TOMMASO MEO†‡

*Institut für Immunologie der Universität München, D-8000 München 2, Federal Republic of Germany; and †Unité d'Immunogénétique and Institut National de la Santé et de la Recherche Médicale U 276, Institut Pasteur, 75724 Paris Cédex 15, France

Communicated by L. L. Cavalli-Sforza, October 15, 1985

ABSTRACT **Antigen HLA-B27 is a high-risk genetic factor with respect to a group of rheumatoid disorders, especially ankylosing spondylitis. A cDNA library was constructed from an autozygous B-cell line expressing HLA-B27, HLA-Cw1, and the previously cloned HLA-A2 antigen. Clones detected with an HLA probe were isolated and sorted into homology groups by differential hybridization and restriction maps. Nucleotide sequencing allowed the unambiguous assignment of cDNAs to *HLA-A*, *-B*, and *-C* loci. The HLA-B27 mRNA has the structural features and the codon variability typical of an HLA class I transcript but it specifies two uncommon amino acid replacements: a cysteine in position 67 and a serine in position 131. The latter substitution may have functional consequences, because it occurs in a conserved region and at a position invariably occupied by a species-specific arginine in humans and lysine in mice. The availability of the complete sequence of HLA-B27 and of the partial sequence of HLA-Cw1 allows the recognition of locus-specific sequence markers, particularly, but not exclusively, in the transmembrane and cytoplasmic domains.**

The HLA-A, -B, and -C (class I) antigens are integral membrane proteins composed of two noncovalently associated molecules: a highly polymorphic glycosylated heavy chain (44 kDa) encoded in the major histocompatibility complex (MHC) on chromosome 6 and the invariant β_2-microglobulin polypeptide (12 kDa) encoded on chromosome 15. These antigens are expressed essentially on all nucleated cells and have the role of restricting recognition by cytotoxic T lymphocytes of tissue allografts and virus-infected cells (for review, see ref. 1). Moreover, the human MHC markers have additional biological and epidemiological relevance, because a number of diseases are known to be associated with HLA (reviewed in ref. 2). Although most of the illnesses previously found associated with class I markers and characterized by immunological overtones have later been resolved into two major groups that correlate more strongly with HLA class II and class III genes, the class I antigen HLA-B27 remains a striking exception (3). The case of this antigen is also very notable owing to the unusually high risk that HLA-B27-positive individuals have of contracting various forms of rheumatoid disorders, among which ankylosing spondylitis (AS) predominates. The strong association of this disease in various ethnic groups with *HLA-B27* [and not with other genes of the HLA complex (3)], the apparent dominant mode of inheritance of AS (4, 5), and the finding that HLA-B27 crossreacts with epitopes found on bacteria implicated in the course of the disease (2) provide the main evidence supporting a direct etiological contribution of this polymorphic marker to the pathogenesis of the disorder. However, other authors (cf. refs. 4, 5) interpret the lack of an absolute epidemiological association as evidence for the existence of a distinct "illness susceptibility" gene and regard the HLA marker as an incidental indicator of the mutant due to locus polymorphism and linkage disequilibrium. We sought to determine the complete sequence of an HLA-B27 mRNA as a direct approach to the primary structure of the antigen and to the nucleotide information necessary for the search at the genomic level of the putative "illness" locus. The choice of this approach was also dictated by an interest in obtaining the molecular tools for testing the basis of the crossreactivity of anti-HLA-B27 antibodies with certain bacterial cell walls.

MATERIALS AND METHODS

Recombinant cDNA clones were prepared from size-selected mRNA isolated from the autozygous B-cell line LG-2 (6) established from individual V.I. of the inbred family STA (7). Synthesis, size selection, and insertion of double-stranded cDNA by dG·dC homopolymer tailing in the *Pst* I site of pBR322 used for transformation of *Escherichia coli* strain C600 were essentially according to standard protocols (8).

RESULTS

Detection and Grouping of HLA Class I-Encoding cDNA Clones. The approach utilized for the isolation of HLA-B27 cDNA clones involved the preparation of a cDNA library from a cell line autozygous for the HLA antigens A2, B27, and Cw1. It was anticipated that the selection of a genotype carrying an *HLA-A* allele whose sequence was known (9) and the availability of the HLA-B7 (10) and the HLA-Cw3 (11) sequences would ease the identification of the sought cDNA. Clones encoding the HLA class I antigens were detected by hybridization with an HLA full-length cDNA probe (10, 12). As studies in the murine (13) and human (14) systems have shown that sequences from the 3′ untranslated part (3′ UT) can be used for the identification of certain individual class I genes, we attempted to group the cDNA clones of the LG-2 cell line by means of two such described probes (14). Of 40 class I-positive clones, 8 gave a stronger signal with the HLA-B-specific probe (no. 2 in Fig. 1) and 12 gave a stronger signal with the HLA-A-specific probe (no. 3 in Fig. 1). Clones displaying weak hybridization signals probably contain insertions covered only partially by the two probes or they represent sequences less closely related to the *HLA-A* and *-B* genes. Indeed, in contrast to the relative homogeneity found within the putative HLA-A and HLA-B cDNA groups (Fig.

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: AS, ankylosing spondylitis; bp, base pair(s); MHC, major histocompatibility complex; 3′ UT, 3′ untranslated part.

‡To whom correspondence should be addressed.

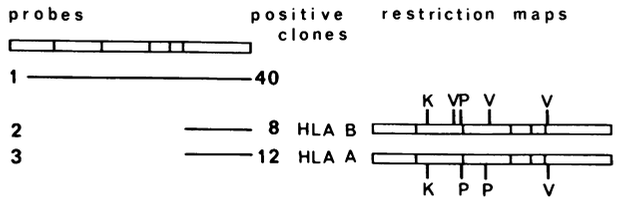

FIG. 1. Detection and locus assortment of HLA class I cDNA clones. Clones were detected by colony hybridization (washes were done at least twice in 0.15 M NaCl/0.015 M trisodium citrate at 65°C) and classified by the probes indicated: probe 1, pHLA-2 (10, 12); probe 2, pHLA-1.1, HLA-B specific (14, 15); probe 3, pHLA 2a.1, HLA-A specific (14). The HLA-A and -B groupings found by differential hybridization were confirmed by the restriction enzyme sites shown: K, *Kpn* I; P, *Pst* I; V, *Pvu* II. The segmentation denotes domains 1–3 and the transmembrane and the cytoplasmic region of the protein, followed by the 3′ UT of the mRNA.

1), a restriction map analysis showed that the third group includes different kinds of transcripts, one of which (clone pC1) attracted our attention for the unexpected presence of sites diagnostic of the HLA-B group.

Identification and Characterization of HLA-B27 and HLA-Cw1 cDNA Clones. Partial sequencing (data not shown) of clones from the putative HLA-A group definitely identified them as transcripts of the *HLA-A2* gene (16) in accordance with the genotype of the cell line. Thus, to detect an HLA-B27-specific cDNA clone we focused on the second group of plasmids (pB1-4) and on clone pC1 (Fig. 2). The nucleotide sequence comparisons reported in Fig. 3 show that clone pC1 carries an HLA-C-related sequence and that the pB set of clones encodes a sequence strongly homologous to HLA-B7. On the basis of the phenotype of the autozygous cell line we conclude that clone pC1 encodes part of the antigen HLA-Cw1, whereas the sequence identified by the series of pB clones specifies the antigen HLA-B27. The nucleotide sequence merged from the two overlapping clones pB1 [1372 base pairs (bp)] and pB3 (605 bp) spans the entire coding region and the 3′ UT end of an HLA-B27 mRNA. The overall features of this transcript are those typical of a class I heavy chain mRNA. Clone pB3 encodes 21 amino acids of the signal peptide and terminates prior to the ATG starting codon, which for the known HLA class I gene sequences is located several base pairs upstream (11, 16, 19). A translation termination codon was found at the end of the fourth exon (corresponding to triplet 274 in Fig. 3) of clone pB1. However, two additional clones, pB2 and pB4, showed a TGG codon in this position. Furthermore, no evidence could be found for the expression of a truncated HLA polypeptide either by *in vitro* translation of hybrid-selected mRNA or by immunoprecipitations of metabolically labeled cell products (data not shown). Considering also that the reverse transcriptase used for the synthesis of both cDNA strands is known to be more error prone than other polymerases, we conclude that this TGA codon is an artefact of clone pB1. The authentic termination codon found in clones pB2 and pB4 as well as pB1 is located at the same position of exon 7 as in HLA-B7 (10). No other nucleotide difference was found among the sequenced insertions of the pB clones.

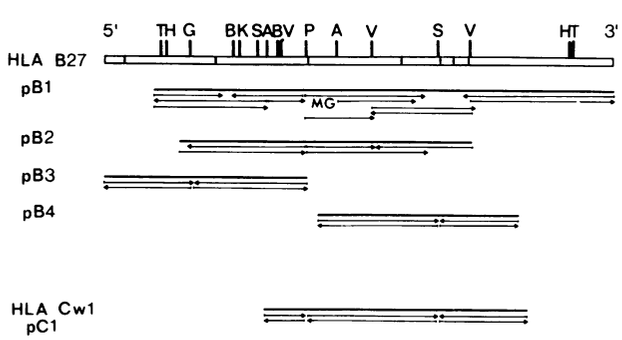

FIG. 2. Sequencing strategies. cDNA fragments shown were subcloned and sequenced by the dideoxy method (17); an ambiguous stretch (MG) in clone pB1 was verified by the Maxam and Gilbert protocol (18). Restriction sites are shown in relation to the exons of a typical HLA heavy chain including the 3′ UT region. A, *Sau*3A; B, *Bgl* I; G, *Bgl* II; H, *Hin*fI; K, *Kpn* I; P, *Pst* I; S, *Sac* I; T, *Taq* I; V, *Pvu* II.

DISCUSSION

The predicted protein sequence of HLA-B27 antigen (Fig. 4) has the known characteristics of an HLA heavy chain (1), with conserved cysteine residues for the disulfide bridges of the second (positions 101 and 164) and the third (positions 203 and 259) domains, and the single N-glycosylation site (position 86). In the two external variable domains of the protein, HLA-B27 displays the same degree of divergence from HLA-B7 as from HLA-B40. Seeking possible structural correlates of the unique features of this antigen it is interesting to consider the position and the nature of the amino acid replacements that distinguish HLA-B27 from other human class I molecules. As expected, most of the amino acid exchanges occur at positions or within segments responsible for the variability of class I antigens (21, 23). Nevertheless, seven unique positions can be recognized where HLA-B27 carries a residue not found in any of the HLA-A, -B, and -C sequences comparable at present (Fig. 4). The replacements at positions 70, 82, 83, 97, and 182 are unremarkable for they either represent conservative substitutions or occupy sites of intra- or interlocus variability. In contrast, two replacements are unusual and deserve special comment. Position 67 is occupied by a potentially reactive cysteine that might influence the intermolecular interactions of HLA-B27 due to its location in an area thought to mediate the binding of monoclonal antibodies as well as of alloantibodies (22). However, it is difficult to assess the relevance of such an amino acid substitution in view of the observation that a free cysteine occupies, for example, a similarly polymorphic site in the HLA-Cw3 sequence (position 9). In contrast, the appearance of a serine in position 131 is likely to have a more profound effect on the conformation of HLA-B27, because this nonconservative replacement occurs at a site invariably occupied by a species-associated arginine residue in human and lysine in mice (22). The species specificity of this residue is confirmed in all murine class I gene sequences so far known, which include *H-2K*^{q}, *H-2K*^{w8} (see ref. 24), *Q 8* and *Q 10* (BALB/c, see ref. 25), and *Q 4–Q 10* (C57BL/10; unpublished results); this position is occupied by an arginine also in HLA-Aw24 (26) as well as in a rabbit class I antigen (27). It is noteworthy that this site is embedded in segment 122–136, which defines an area submitted to stringent selective constraints because it represents the evolutionarily most conserved region in the second domain of class I molecules. As no secondary structure could be predicted for the segment 129–136 (28), it is impossible to guess what consequence the replacement of a bulky polar residue by serine would have on the tertiary structure of the antigen.

Our approach to clone HLA-B27 depended on the verification of the postulate that the *HLA-A*, *-B*, and *-C* allelic series can indeed be recognized at the nucleotide and/or protein level. In particular, the isolation of clone pC1 encoding a protein whose incomplete sequence is essentially superimposable to that deduced from *HLA-Cw3*, the only *HLA-C* allele so far sequenced, allows us to make some preliminary generalizations on the sequence markers diagnostic of each HLA locus. One such locus-specific trait

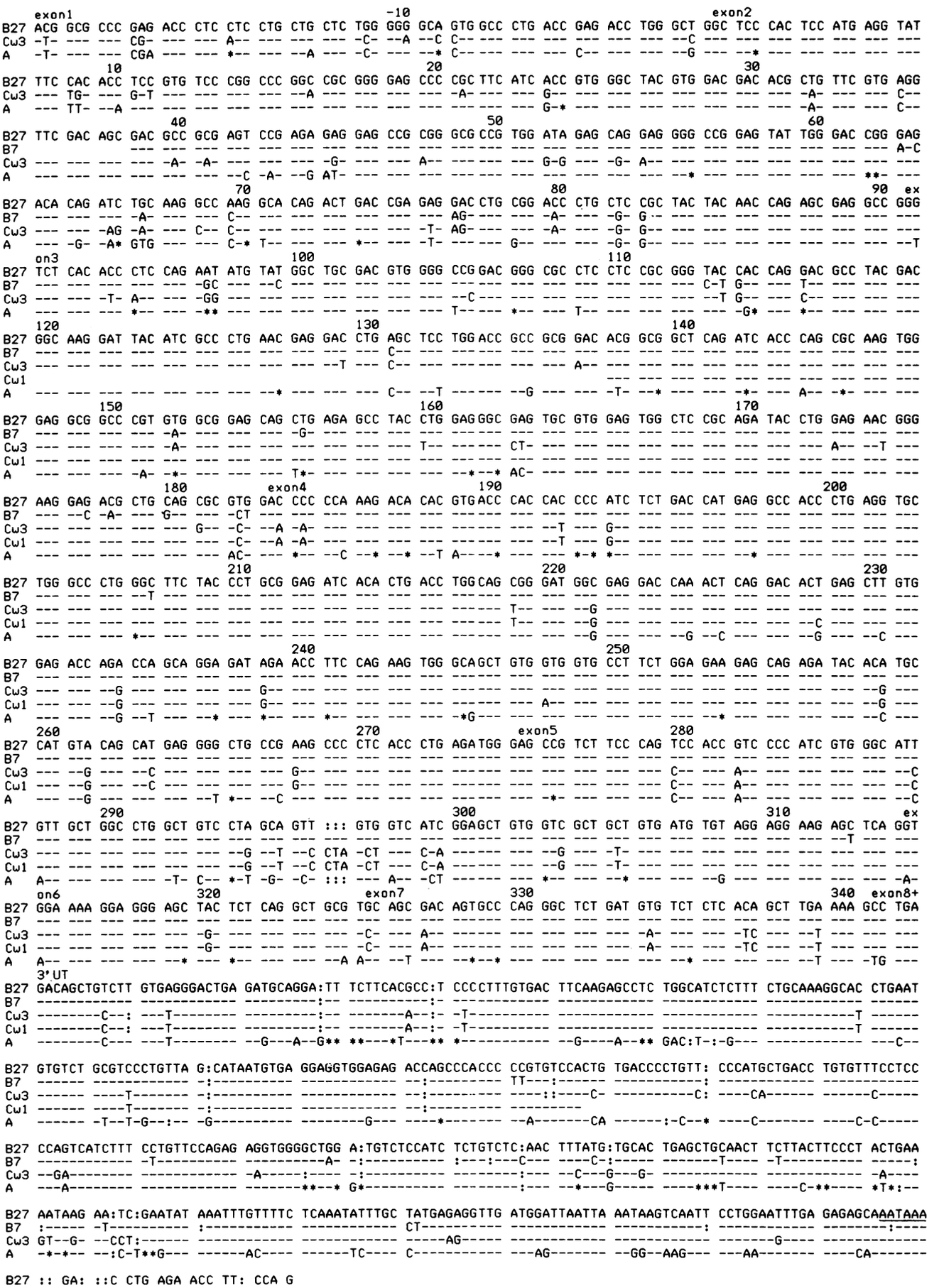

FIG. 3. (*Legend appears on the opposite page.*)

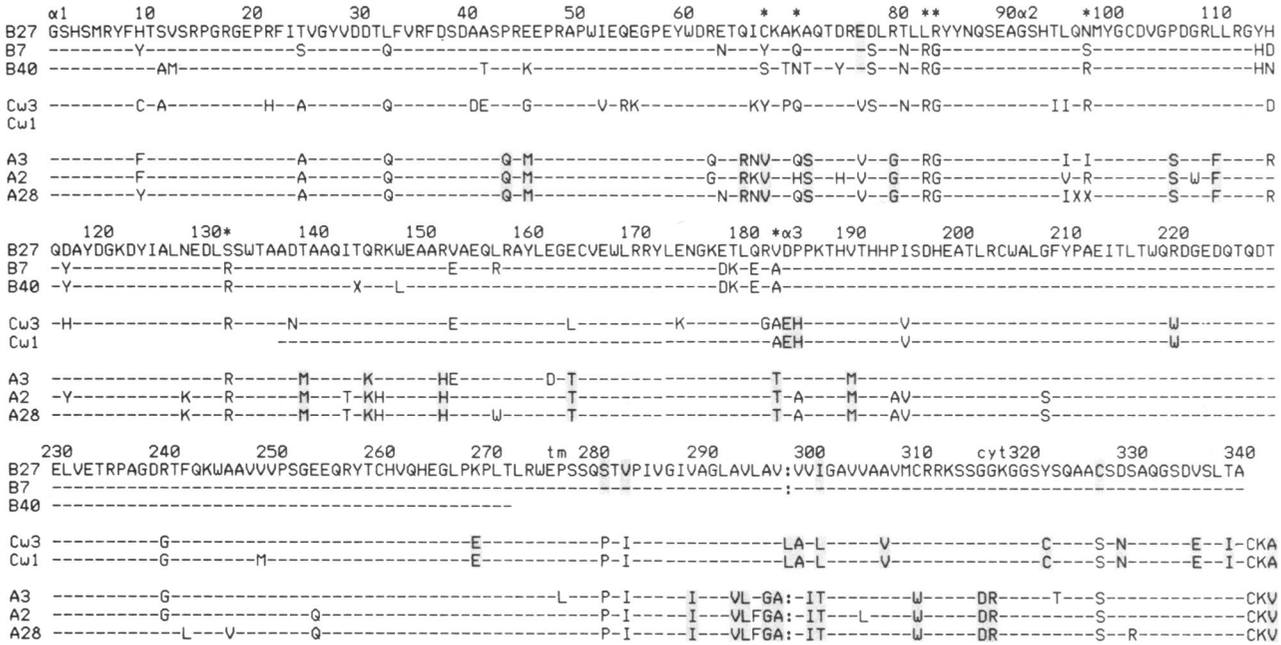

FIG. 4. Alignment of the amino acid sequences predicted for HLA-B27 and -Cw1 with other HLA proteins. Partial sequences from nucleotide data were completed with residues taken from protein data: B7 (21); B40 (22); A28 (21). Dashes indicate identity, asterisks mark positions discussed in the text, and colons fill in gaps introduced to maximize matches. The one-letter code for amino acids is used. Residues distinctly associated with each class I locus are shadowed.

resides in the different lengths of each mature HLA-A, -B, and -C protein, which appear to be 341, 338, and 342 amino acids long, respectively. Furthermore, several residues can be easily discerned for their characteristic conservation within each of the three sets of polypeptides (shadowed in Fig. 4) or nucleotide sequences. It is noteworthy that these diagnostic residues occur at positions differing among the three allelic series. Evidently, they mark regions not affected by recent events of intrachromosomal sequence exchanges, of the type postulated to contribute to the generation of the extraordinary polymorphism of class I genes (29). The number of marker residues is notably higher in HLA-A and HLA-C than in the HLA-B series. The HLA-B character of the latter group is at the protein level practically confined to the transmembrane and cytoplasmatic regions. This lack of "B-ness" in the $\alpha 1$ and $\alpha 2$ domains is striking in comparison to the 13 positions that in the same region are characteristic of the HLA-A group. The discrepancy is unlikely due to a biased sampling of serologically related sequences since both sets of alleles include crossreactive antigens. It seems more plausible that in this part of the molecule the HLA-B character has been eroded by an extraordinary high level of polymorphism combined with the clustering of variability within the $\alpha 1$ and $\alpha 2$ domains of class I antigens (21, 23). If additional data will bring to light intralocus subsets of sequences more related to each other, this finding will strengthen the suspicion that some of the presently defined loci actually include pseudoallelic series (30).

A search for interlocus homologies shows that in the 3′ UT region the two HLA-C sequences share many nucleotide matches with the HLA-B sequences at positions where the HLA-As have instead a different and usually also invariant residue (see Fig. 3). This high degree and pattern of similarity is also evident at the amino acid level (7 matches over 39 positions are specifically common to the B/C sequences). It will be interesting to extend the comparisons to introns and flanking sequences to verify the indication that the *HLA-B* and *-C* loci originated through a more recent gene duplication.

On the other hand, the sequence markers unique to *HLA-A*, *-B*, and *-C* do not coincide with those specific for the mouse *H-2K*, *-D*, and *-L* (13), thus providing strong evidence that, in analogy with certain class III loci (31), the duplication of class I genes occurred after the separation of the ancestors of the two species. This conclusion is in accord with the strong intraspecies conservation found in the introns of class I genes (see refs. 16 and 25) as well as with the finding of an amino acid gap in the region encoded by exon 7, which is common to *H-2K*, *D*, and *L* and to the class I-related *Q* locus but is not present in any of the human class I sequences (32).

In conclusion, HLA-B27 features two unique amino acid substitutions that may critically affect the intermolecular interactions of this antigen. It appears now crucial to determine whether these changes are common to the HLA-B27 subtypes defined by serological and biochemical (33, 34) studies as well as by T-lymphocyte assays (35, 36), neither of which shows a preferential correlation with AS. The finding that these substitutions are conserved among the subtypes would strengthen the relevance of these amino acids for the major antigenic epitope(s) of HLA-B27 and as a molecular site of correlation with the disease.

Alternatively, if a putative illness susceptibility gene in linkage disequilibrium with *HLA-B27* (with a coefficient of 0.019 according to ref. 4—i.e., about 90% of its possible

FIG. 3. Nucleotide sequence alignments of HLA-B27 and -Cw1 in comparison with already reported HLA sequences. B7, partial cDNA sequence (10); Cw3, genomic sequence (11); A, constructed by comparing the nucleotide sequences of three *HLA-A* genes: *A3* (19), *A2* (16), and *A28* (20); only matches common to the three sequences are reported, whereas variable positions are marked by asterisks. Nucleotides are shown in triplets for coding regions and in blocks of 12 for the 3′ UT. Dashes are used to indicate identity and colons are used to fill in gaps introduced to maximize matches. The poly(A) signal is underlined. Numbered sites refer to codon positions of the processed polypeptide. Note that the presence of an extra codon (296/297) in HLA-C sequences coincides with the direct repeat of the nonanucleotide GCTGTCCTA.

maximum) is responsible for the diseases associated with this antigen (4, 5), the appearance of this mutation must have predated the emergence of the *HLA-B27* subtypes, as none of them displays a stronger linkage disequilibrium with the postulated "illness gene." If the recombination distance (r) between the hypothetical gene and *HLA-B27* were $\geq 10^{-3}$, this linkage disequilibrium should have practically vanished because of the decrement expected under Hardy–Weinberg conditions of $(1 - r)$ per generation and because the mutant must be at least 25,000–100,000 years old—the time (i.e., 1250–5000 generations) elapsed since the divergence of the major human races (37). Furthermore, the frequency of 0.022 estimated for the AS susceptibility gene (4) suggests that, according to the theoretical results obtained by Kimura and Ohta (38), the mutant may be older and even as old as the life span of the human species. Therefore, assuming an average of 10^6 bp of DNA per centimorgan (39), if such a susceptibility gene does indeed exist, it should lie at most within a few thousand base pairs from *HLA-B27*. This study has provided the molecular basis for a direct test of the bacterial cross-tolerance hypothesis for the HLA-B27 pathogenicity as well as the structural data that can expedite the finding of the postulated susceptibility locus contiguous to the *HLA-B27* gene.

After this work was completed, Ezquerra *et al.* (40) reported the sequence of a papain-solubilized *HLA-B27* antigen. Their work conveys information on the primary structure of the extracellular domains of the molecule, which, except for an irrelevant change at position 182, is identical to the one here deduced from the nucleotide sequence.

We thank W. Leibold for the LG-2 cell line, H. Grosse-Wilde for HLA typing, H. T. Orr, H. L. Ploegh, and S. M. Weissman for DNA probes, B. Jordan and J. Lopez de Castro for exchanging information before publication, and S. Heuser, D. Schendel, J. Johnson, and M. Tosi for comments on the manuscript. This study was supported in part by the Deutsche Forschungsgemeinschaft (SFB 37/217) and Genzentrum at the University of Munich.

1. Ploegh, H. L., Orr, H. T. & Strominger, J. L. (1981) *Cell* **24**, 287–299.
2. Möller, G., ed. (1983) *Immunol. Rev.* **70**.
3. Svejgaard, A., Friis, J., Morling, N. & Ryder, L. P. (1981) *J. Rheumatol.* **8**, 541–548.
4. Kidd, K. K., Bernoco, D., Carbonara, A. O., Daneo, V., Steiger, U. & Ceppellini, R. (1977) in *HLA and Disease*, eds. Dausset, J. & Svejgaard, A. (Munksgaard, Copenhagen), pp. 72–80.
5. Thomson, G. & Bodmer, W. F. (1977) in *HLA and Disease*, eds. Dausset, J. & Svejgaard, A. (Munksgaard, Copenhagen), pp. 84–93.
6. Gatti, R. A. & Leibold, W. (1979) *Tissue Antigens* **13**, 35–44.
7. Joergensen, F., Lamm, U. L. & Kissmeyer-Nielssen, F. (1973) *Tissue Antigens* **3**, 323–339.
8. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).
9. Lopez de Castro, J. A., Strominger, J. L., Strong, D. M. & Orr, H. T. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 3813–3817.
10. Biro, P. A., Pan, J., Sood, A. K., Kole, R., Reddy, V. B. & Weissman, S. M. (1983) *Cold Spring Harbor Symp. Quant. Biol.* **47**, 1079–1086.
11. Sodoyer, R., Damotte, M., Delovitch, T. L., Jordan, B. R. & Strachan, T. (1984) *EMBO J.* **3**, 879–885.
12. Sood, A. K., Pereira, D. & Weissman, S. M. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 616–620.
13. Kress, M., Liu, W. Y., Jay, E., Khoury, G. & Jay, G. (1983) *J. Biol. Chem.* **258**, 13929–13936.
14. Koller, B. H., Sidwell, B., DeMars, R. & Orr, H. T. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 5175–5178.
15. Ploegh, H. L., Orr, H. T. & Strominger, J. L. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 6081–6085.
16. Koller, B. H. & Orr, H. T. (1985) *J. Immunol.* **134**, 2727–2733.
17. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
18. Maxam, A. M. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 560–564.
19. Strachan, T., Sodoyer, R., Damotte, M. & Jordan, B. R. (1984) *EMBO J.* **3**, 887–894.
20. Arnot, D., Lillie, J. W., Auffrey, C., Kappes, D. & Strominger, J. L. (1984) *Immunogenetics* **20**, 237–252.
21. Orr, H. T., Lopez de Castro, J. A., Parham, P., Ploegh, H. L. & Strominger, J. L. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 4395–4399.
22. Lopez de Castro, J. A., Bragado, R., Strong, D. M. & Strominger, J. L. (1983) *Biochemistry* **22**, 3961–3969.
23. Maloy, W. L. & Coligan, J. E. (1982) *Immunogenetics* **16**, 11–25.
24. Morita, T., Delarbre, C., Kress, M., Kourilsky, P. & Gachelin, G. (1985) *Immunogenetics* **21**, 367–383.
25. Sher, B. T., Nairn, R., Coligan, J. E. & Hood, L. E. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 1175–1179.
26. N'Guyen, C., Sodoyer, R., Trucy, J., Strachan, T. & Jordan, B. R. (1985) *Immunogenetics* **21**, 479–489.
27. Marche, P. N., Tykocinski, M. L., Max, E. E. & Kindt, T. J. (1985) *Immunogenetics* **21**, 71–82.
28. Vega, M. A., Bragado, R., Ezquerra, A. & Lopez de Castro, J. A. (1984) *Biochemistry* **23**, 823–831.
29. Brégégère, F. (1983) *Biochimie* **65**, 229–237.
30. Bodmer, W. F. (1972) *Nature (London)* **237**, 139–145.
31. Lévi-Strauss, M., Tosi, M., Steinmetz, M., Klein, J. & Meo, T. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 1746–1750.
32. Guild, B. C. & Strominger, J. L. (1984) *J. Biol. Chem.* **259**, 9235–9240.
33. Mölders, H. H., Breuning, M. H., Ivanyi, P. & Ploegh, H. L. (1983) *Human Immunol.* **6**, 111–117.
34. Choo, S. Y., Antonelli, P., Nepom, G. & Hansen, J. A. (1985) *Fed. Proc. Fed. Am. Soc. Exp. Biol.* **44**, 557 (abstr.).
35. Breuning, M. H., Lucas, C. J., Breur, B. S., Engelsma, M. Y., de Lange, G. G., Dekker, A. J., Biddison, W. E. & Ivanyi, P. (1982) *Hum. Immunol.* **5**, 259–268.
36. Toubert, A., Gomard, E., Grumet, F. C., Amor, B., Muller, Y. & Lévy, J.-P. (1984) *Immunogenetics* **20**, 513–525.
37. Cavalli-Sforza, L. L. (1969) in *Proc. 12th Int. Cong. Genet. Tokyo* (Science Council of Japan), Vol. 3, pp. 405–416.
38. Kimura, M. & Ohta, T. (1973) *Genetics* **75**, 199–212.
39. Kornit, D. M. (1979) *Lancet* **i**, 104.
40. Ezquerra, A., Bragado, R., Vega, M. A., Strominger, J. L., Woody, J. & Lopez de Castro, J. A. (1985) *Biochemistry* **24**, 1733–1741.