Proc. Natl. Acad. Sci. USA
Vol. 87, pp. 2833–2837, April 1990
Immunology

Rapid cloning of HLA-A,B cDNA by using the polymerase chain reaction: Frequency and nature of errors produced in amplification

(polymorphism/gene families/recombination/histocompatibility)

PETER D. ENNIS, JACQUELINE ZEMMOUR, RUSSELL D. SALTER, AND PETER PARHAM*

Department of Cell Biology, Stanford University, Stanford, CA 94305

Communicated by D. Bernard Amos, January 2, 1990

ABSTRACT **A method for cloning full-length HLA-A,B cDNA (1.1 kilobases) by using the polymerase chain reaction (PCR) is described. Six *HLA-A,B* alleles (*HLA-A2*, *-A25*, *-B7*, *-B37*, *-B51*, and *-B57*) were cloned, and their structures were determined. Multiple PCR clones for each allele were sequenced to obtain both an accurate consensus sequence and an "authentic" clone having that sequence. Sequences from 50 clones encoding five different alleles permit assessment of the frequency and nature of PCR-produced errors. These include recombinations, deletions, and insertions in addition to point substitutions. Authentic clones were obtained at a frequency of between 30% and 70%, and analysis of three or four clones generally should be sufficient for characterization of an allele.**

HLA-A,B,C genes encode cell surface glycoproteins that, in association with β_2-microglobulin and a peptide, form the ligand for the antigen receptor of $CD8^+$ T lymphocytes (1, 2). They are probably the most polymorphic of human genes, their diversity resulting from positive evolutionary selection and correlating with the capacity to respond to different antigens (3–5). Although structures for over 50 *HLA-A,B,C* alleles have been determined, many others remain uncharacterized (6). Current methods for analysis of *HLA-A,B,C* alleles are limited by the time and effort required to make and screen either genomic or cDNA libraries. The polymerase chain reaction (PCR) provides an approach that could eliminate these procedures and potentially speed the acquisition of allelic sequences (7). In this paper we describe the application of PCR to the cloning and sequence determination of class I HLA cDNA.†

MATERIALS AND METHODS

Cell Lines. The following human Epstein–Barr virus-transformed cell lines were used in this study: JY (*HLA-A2*, *HLA-B7*); BM92 (*HLA-A25*, *HLA-B51*); WIN (*HLA-A1*, *HLA-B57*); MOC (*HLA-A1*, *HLA-A2*, *HLA-B13*, *HLA-B57*); KASO (*HLA-A1*, *HLA-B37*); and MG (*HLA-A1*, *HLA-A30*, *HLA-B13*, *HLA-B37*). The boldface genes are those for which cDNAs were cloned. Cells were grown in RPMI 1640 medium containing 10% (vol/vol) fetal calf serum and supplemented with glutamine, penicillin, and streptomycin.

Preparation and Amplification of cDNA. One hundred-milliliter cultures of cells were used to prepare total cellular RNA (8). First-strand cDNA was synthesized by using oligo(dT) and avian myeloblastosis virus reverse transcriptase (9) and was trace-labeled with [α-^{32}P]dCTP. Approximately 1% of total radioactivity was incorporated into a trichloroacetic acid-precipitable form. Half of the product was extracted and back extracted with 1:1 (vol/vol) phenol/chloroform, twice extracted with ether, and precipitated with ethanol. The precipitate was redissolved in 50 μl of water, and 3 μl was used as the target for PCR amplification, which was performed by using GeneAmp kits and a DNA thermal cycler (Perkin–Elmer/Cetus). The reaction was in 100 μl and used 50 pmol of each primer. Two protocols were used: amplifications using protocol 1 went for 30 cycles in which each cycle consisted of 60 sec at 94°C, 60 sec at 65°C, and 90 sec at 72°C. Experiments using protocol 2 used 20 cycles with 60 sec at 94°C, 1 sec at 65°C, and a variable time at 72°C, which started at 50 sec and increased incrementally by 1 sec in each cycle. In both protocols the amplification finished with 10 min at 72°C.

Subcloning and Sequencing. Half of the PCR product was extracted and back-extracted with phenol/chloroform, precipitated with ethanol, and digested with 40 units of *Hin*dIII for 1 hr at 37°C. After similar extraction and precipitation, the *Hin*dIII-cut product was digested with 40 units of *Sal* I for 1 hr at 37°C. Double-cut product was purified with glass beads and ligated to similarly cut M13mp18 and M13mp19 vectors, which were used to transform competent JM109 cells (10). Phage were picked, grown in liquid culture, and used to prepare single-stranded M13 DNA, which was then sequenced, and the sequences were analyzed as described (11).

RESULTS

Our goal was to isolate and sequence cDNA clones encoding complete class I HLA heavy chains. Total RNA from human B-cell lines was used to prepare single-stranded cDNA, and this provided the substrate for specific amplification of *HLA-A,B,C* sequences by the PCR. The oligonucleotides used to prime specific amplification derive from relatively conserved sequences in the 5′ and 3′ untranslated regions of *HLA-A,B,C* genes (Fig. 1). Restriction sites for *Sal* I and *Hin*dIII were incorporated into the 5′ and 3′ primers, respectively, to enable the amplification products to be subcloned into sequencing and expression vectors. A four-base spacer sequence was placed external to the restriction site, a preliminary experiment having shown that amplification products without a spacer could not be cut efficiently with either enzyme.

Amplification resulted in the expected product of ≈1.1 kilobases (kb), which was quite pure as assessed by electrophoresis (Fig. 2). This product, predicted to contain a mixture of *HLA-A,B,C* sequences, was directionally subcloned into M13mp18 and M13mp19 vectors and, after transformation of *Escherichia coli*, individual recombinant phage were picked

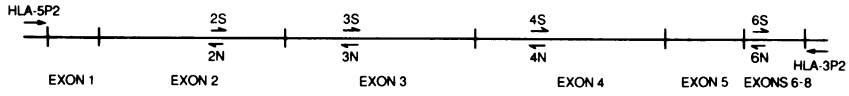
The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: PCR, polymerase chain reaction.

*To whom reprint requests should be addressed.

†The sequences for *HLA-A25*, *-A2*, *-B7*, *-B57*, *-B51*, and *-B37* reported in this paper have been deposited in the GenBank data base (accession nos. M32321, M32322, M32317, M32318, M32319 and M32320, respectively).

A



B

priming oligonucleotides		5'	3'
HLA-5P2	5' untranslated region	GGGCGTCGACGGACTCAGAATCTCCCCAGACGCCGAG	
HLA-3P2	3' untranslated region	GCCCAAGCTTTCTCAGTCCCTCACAAGGCAGCTGTC	
sequencing oligonucleotides			
2S	exon 2 236-255	AGGGGCCGGAGTATTGGGAC	
2N	exon 2 236-255	GTCCCAATACTCCGGCCCCT	
3S	exon 3 429-450	CGGCAAGGATTACATCGCCCTG	
3N	exon 3 429-450	CAGGGCGATGTAATCCTTGCCG	
4S	exon 4 703-724	GCGGAGATCACACTGACCTGGC	
4N	exon 4 703-724	GCCAGGTCAGTGTGATCTCCGC	
6S	exon 6 1023-1043	AGGGAGCTACTCTCAGGCTGC	
6N	exon 6 1023-1042	TGCAGCCTGAGAGTAGCTCCCT	

FIG. 1. (*A*) Schematic of a HLA class I cDNA PCR product showing PCR primers (large arrows), sequencing primers (small arrows), and exon boundaries (vertical lines). Arrowheads are at the 3' hydroxyl end of each oligonucleotide primer and point in the direction of polymerase extension. (*B*) Sequences of priming and sequencing oligonucleotides. Priming oligonucleotides were designed from comparison of the 5' and 3' untranslated regions of *HLA-A2.1* (12), *-A3* (13), *-A24* (3), *-B44* (14), *-Bw58* (15), *-Cw1*, *-Cw2* (16), and *-Cw3* (17) with selection of conserved sequences. The *Sal* I site in *HLA-5P2* and the *Hin*dIII site in *HLA-3P2* are underlined. Sequencing oligonucleotides S are derived from the sense strand; and N, from the antisense strand.

and analyzed by limited sequencing with one or two oligonucleotide primers (Fig. 1). We now have analyzed well over 100 clones, and without exception they all contained class I *HLA* inserts. On the basis of locus-specific nucleotides (5), these clones can, by sequencing with oligonucleotides 3S and 4N, be assigned to an *HLA* locus. Although one or two clones from *HLA-C* and other class I genes have been tentatively identified, almost all of the clones obtained derive from *HLA-A* and *HLA-B*, and analysis was concentrated on them.

We anticipated that replication errors introduced during the PCR would necessitate sequencing of multiple clones to determine the true sequence of an allele and to identify a clone having that sequence. To assess the magnitude of this problem, we studied five alleles from three cell lines: *HLA-A2* and *HLA-B7* from JY, *HLA-A25* and *HLA-B51* from BM92, and *HLA-B57* from WIN. For each allele the sequences of 10 clones—five from the sense strand and five from the antisense strand—were determined. In each case a clear consensus sequence was obtained (Fig. 3), and a significant number of the clones (four for *HLA-A2*, three for *HLA-B7*, six for *HLA-A25*, three for *HLA-B51*, and seven for *HLA-B57*) had sequences that were identical to the relevant consensus and most probably represent faithful copies of the gene (Table 1). These results clearly show the feasibility of this approach for the isolation and characterization of *HLA-A,B* alleles.

Between 3 and 7 clones in each set of 10 had one or more mutations with respect to the consensus sequence, and these are listed in Table 1. They are presumably the result of "errors" in the PCR amplification. Every mutation was unique to a single clone, indicating that errors are infrequent events, mostly occurring late in amplification as the target sequences accumulate. Clones having point substitutions, deletions, insertions, and recombinations were all found, giving a total of 37 mutations, including 28 point substitutions, in the 52,579 nucleotides sequenced. The frequency with which errors were found was 1 per 1421 nucleotides.

The amplification of *HLA-A2* and *HLA-B7* and other class I sequences from the JY cell line was performed first and used protocol 1 as described in *Materials and Methods*. Four of the clones studied proved to be recombinants between *HLA-A2* and *HLA-B7* (Table 1). These apparent recombinants presumably arise from premature termination of the polymerase in one cycle with hybridization of the unfinished product to a heterologous strand, followed by extension and completion in a subsequent cycle. Although the frequency of such recombination did not pervert interpretation of the data from JY, a homozygous cell line, its potential for confusing the analysis of heterozygous cells is considerable. Amplification protocol 2 was therefore designed with the goal of reducing these events. The major changes were to reduce the hybridization time from 60 sec to 1 sec, thus favoring the hybridization of short oligonucleotides over longer incomplete amplification products, and to decrease the number of amplification cycles. The experiments with the BM92 and WIN cell lines used protocol 2, and the absence of any recombinants in the clones analyzed suggests that these modifications were effective. The number of point mutations accumulated was also reduced, presumably due to the fewer cycles of amplification (Table 1).

These experiments show the need for analysis of multiple PCR clones but also indicate that the requisite number may generally be less than 10. An experiment was next designed to specifically isolate clones encoding HLA-B37 from two cell lines. The KASO cell line is homozygous for *HLA-B37*, and four *HLA-B* locus clones from that line gave an identical sequence that is distinct from other *HLA-B* alleles. The heterozygous MG cell line expresses both *HLA-B13* and *-B37*. Knowledge of the previously sequenced *HLA-B13* allele allowed us to assign, on the basis of preliminary sequence, three clones as potentially containing *HLA-B37*, and these were completely sequenced. The three clones gave an unambiguous consensus sequence that was identical to that of the *HLA-B37* clones from the KASO cell line. One of the MG *HLA-B37* clones had a sequence identical to that of the consensus, while the other two each had single and distinct point substitutions. Thus, in this experiment we were able to define the sequence of *HLA-B37* and obtain an authentic clone on the basis of four clones from one amplification and three from the other. In addition, the coding regions of the *HLA-B37* genes in the two cells were shown to be identical.

We find the nucleotide sequence encoding HLA-A2 of the JY cell line to be identical to that obtained from the LCL-721 cell line (12), which was used to interpret the crystallographic structure of the JY protein (1, 18), and also to be identical to that from the GM637 cell line (19). *HLA-B51* from the BM92

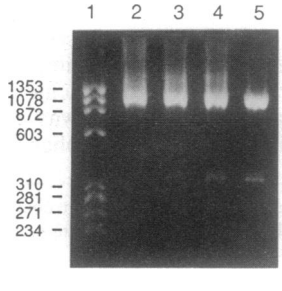


FIG. 2. Agarose gel of 1/10th of the amplification product produced with protocol 1 from the following cell lines: WIN (lane 2), BM92 (lane 3), and MOC (lanes 4 and 5). In lane 1 are markers derived from a *Hae* III digest of ϕX174 replicative form DNA (New England Biolabs).

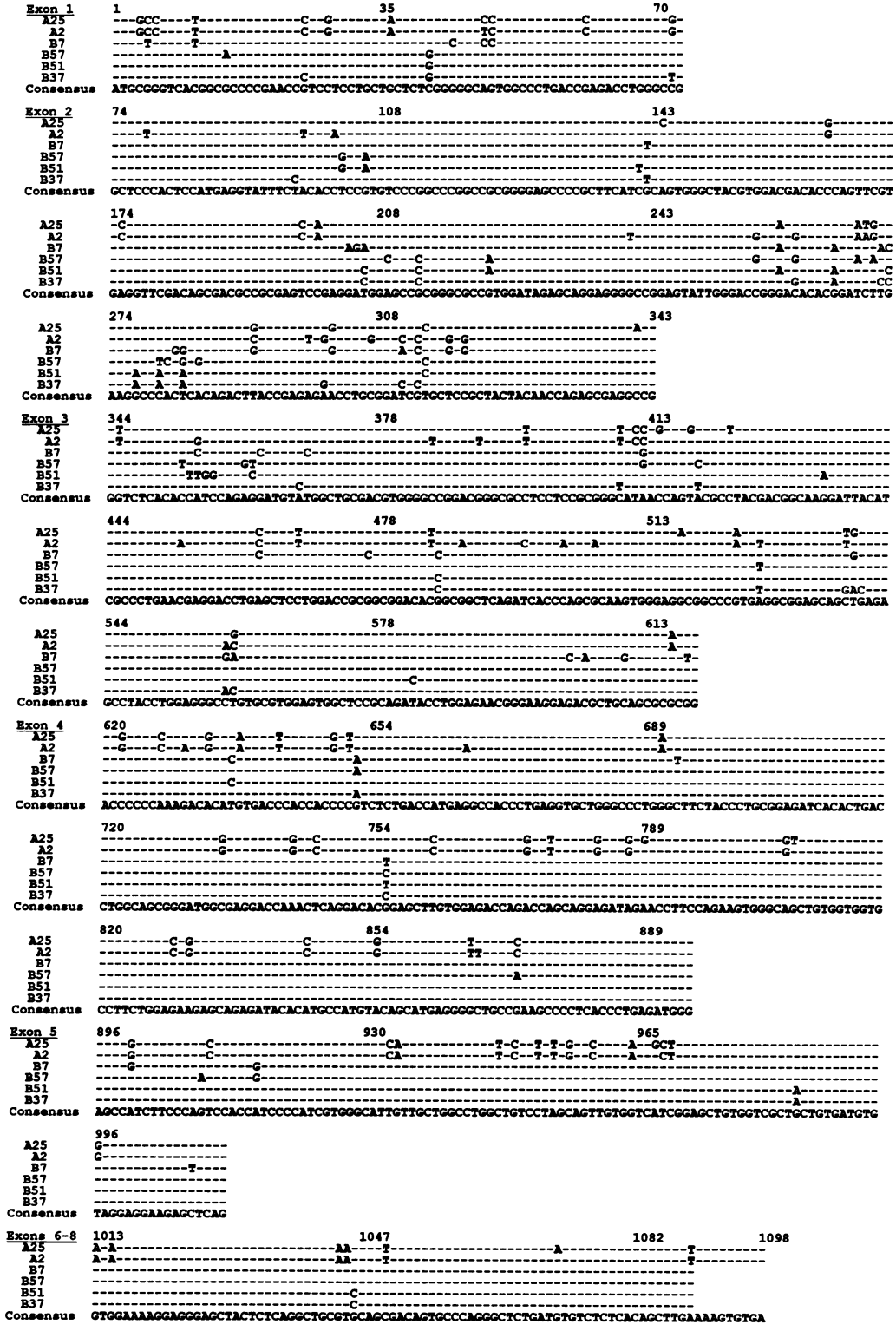


FIG. 3. Consensus nucleotide sequences for the six *HLA-A,B* alleles cloned and sequenced by using PCR amplification. All consensus sequences derive from sequences of 10 individual clones except for *HLA-B37*, which was derived from 7.

cell line was also found to be identical in sequence to *HLA-B51* from two other cell lines (20). In contrast the PCR sequence of HLA-B7 from JY differs from the HLA-B7 cDNA sequence of Sood *et al.* (21) at 26 positions, although the predicted proteins differ by just one amino acid. Comparison of nucleotide sequences from more than 30 *HLA-B* alleles shows that the HLA-B7 cDNA has an unusually high number of silent substitutions, whereas the HLA-B7 PCR sequence does not. This suggests that many of the differences between the PCR and the cDNA sequences are artefacts of the cDNA analysis (21). Our analysis of 10 PCR HLA-B7 cDNA sequences does not support the suggestion that JY expresses two distinguishable *HLA-B7* alleles (22).

HLA-A25, *HLA-B37*, and *HLA-B57* are alleles that have not been sequenced previously. HLA-A10 is serologically split into HLA-A25 and HLA-A26 subtypes. The *HLA-A25* sequence differs from the 8/16 clone of Cianetti *et al.* (19) that probably encodes HLA-A26 by a localized cluster of eight

Table 1. Sequences of clones

Cell line	*HLA* gene	Clone no.	Sequenced nucleotides	Differences from consensus*
JY	*A2*	1	1098	G-685 → A
		2	1098	None
		6	1098	None
		8	1098	None
		9	1098	None
		20	1098	A-361 → G; recombination with *B7*, 561–602; A-652 → G
		23	1098	A-310 → G; T-538 → C; T-965 → C
		25	1098	Recombination with *B7*, 917–932; T-1050 → C
		28	1098	C-117 → T
		29	1098	A-293 → G
JY	*B7*	3	1089	None
		4	1089	None
		7	1089	A-188 → G
		10	1089	C-218 → T
		12	1089	C-111 → T; recombination with *A2*, 874–909
		21	1089	Recombination with A2, 653–734; G-996 → A
		22	1089	C-914 → T
		24	1088	Deletion, 278
		26	1089	None
		27	1089	A-785 → G
BM92	*A25*	201	1098	None
		202	1098	None
		204	1098	C-358 → T
		207	490	None
		208	728	None
		115	1098	G-456 → A
		116	1098	None
		117	1098	None
		118	1098	C-426 → T
		119	1098	T-67 → C; A-908 → G
BM92	*B51*	203	989	None
		205	972	G-126 → A; deletion, 896–1012
		206	1089	None
		211	1041	T-1033 → C
		212	1089	A-839 → G
		101	1089	None
		102	1089	A-437 → G; T-1073 → C
		103	1089+	Insertion of 104 bases, 895
		104	1017	Deletion, 74–145
		105	1089	Deletion, 1039
WIN	*B57*	407	1089	None
		408	1089	None
		411	1089	None
		412	922	G-190 → A
		415	1089	None
		303	1089	None
		304	1089	G-1051 → A
		305	780	None
		307	1089	T-984 → C
		310	830	None

*Gives position and nature of the mutation.

substitutions giving rise to seven amino acid substitutions in the long α-helix of the α_1 domain. Thus, these two alleles are probably related in evolution by a single segmental exchange. Such a simple relationship is not seen between HLA-B57 and its serologically related partner, HLA-Bw58. Although the *HLA-B57* and *-Bw58* alleles clearly share a common ancestor, they differ by 16 nucleotide substitutions that are scattered throughout the coding region and produce eight amino acid differences. HLA-B37 shows relationships with HLA-B18 (residues 1–62) and HLA-Bw47 (residues 63–90) in the α_1 domain and is a composite of segments shared with at least five other HLA-B molecules in the α_2 domain (Fig. 4).

DISCUSSION

Application of molecular biology to the major histocompatibility complex (MHC) has provided the analytical resolution that the complexity of this system demands. In turn, analysis of MHC polymorphism provides a critical test for methods to clone and sequence genes. *HLA-A,B,C* alleles comprise a large family of highly related sequences in which variation is produced by relatively small differences in large numbers of nucleotides, and every difference is potentially important. For example, a single substitution in exon 4 of *HLA-Aw68* results in the protein being unable to bind to the CD8 glycoprotein of T cells (26). We find that a variety of mutational events—recombinations, deletions, insertions, and point substitutions—occur with detectable frequency in amplification with the PCR and that analysis of multiple clones is essential to obtain reliable data. In particular, the identification of recombinations required considerable prior knowledge of the nature of polymorphism in class I *HLA* genes, and such artefacts pose potential complications in the analysis of less-well-characterized genes or DNA sequences for which the extent of polymorphism is not known. By analysis and comparison of multiple PCR clones, we have obtained unambiguous sequences for six *HLA-A,B* alleles and in each case have obtained one or more clones with the authentic consensus sequence. These clones have the potential to be used for expression studies and thereby for immunological analysis of the encoded proteins.

Advantages over traditional cloning approaches are (*i*) the considerable saving of time and effort in the procedures, leading to the isolation of clones; and (*ii*) the predictability or reliability of the approach, in that clones are always full-length and no failures to clone targeted genes have so far occurred. The major disadvantage is the increased sequencing that results from the analysis of multiple clones. However, we judge this to be significantly less burdensome than the making and screening of libraries and the subsequent purification of clones. There has been variability in the numbers and nature of PCR errors in sets of clones from different alleles, making it difficult to gauge a minimum number of clones that must be analyzed to obtain either an accurate sequence or an authentic clone. However, numbers between three and six should generally suffice, and there has been no ambiguity in knowing when an allele is "done." A possible refinement to this approach is the use of locus-specific amplification primers, and this may be necessary for investigation of *HLA-C* and nonclassical class I genes (27).

We thank H. Erlich and R. Higuchi for their help with the PCR and Patricia Massard for preparation of the manuscript. This research was supported by Grant AI24258 from the U.S. Public Health Service. P.P. is a scholar of the Leukemia Society.

1. Bjorkman, P. J., Saper, M. A., Samraoui, B., Bennett, W. S., Strominger, J. L. & Wiley, D. C. (1987) *Nature (London)* **329**, 512–518.
2. Townsend, A. & Bodmer, H. (1989) *Annu. Rev. Immunol.* **7**, 601–624.
3. N'Guyen, C., Sodoyer, R., Trucy, J., Strachan, T. & Jordan, B. R. (1985) *Immunogenetics* **21**, 479–489.
4. Hughes, A. L. & Nei, M. (1988) *Nature (London)* **335**, 167–170.
5. Parham, P., Lawlor, D. A., Lomen, C. E. & Ennis, P. D. (1989) *J. Immunol.* **142**, 3937–3950.
6. Dupont, B., ed. (1989) *The Immunobiology of HLA* (Springer, New York), Vol. 1.
7. Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B. & Erlich, H. A. (1988) *Science* **239**, 487–491.
8. Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J. & Rutter, W. J. (1979) *Biochemistry* **18**, 5294–5299.
9. Okayama, H. & Berg, P. (1982) *Mol. Cell. Biol.* **2**, 161–170.

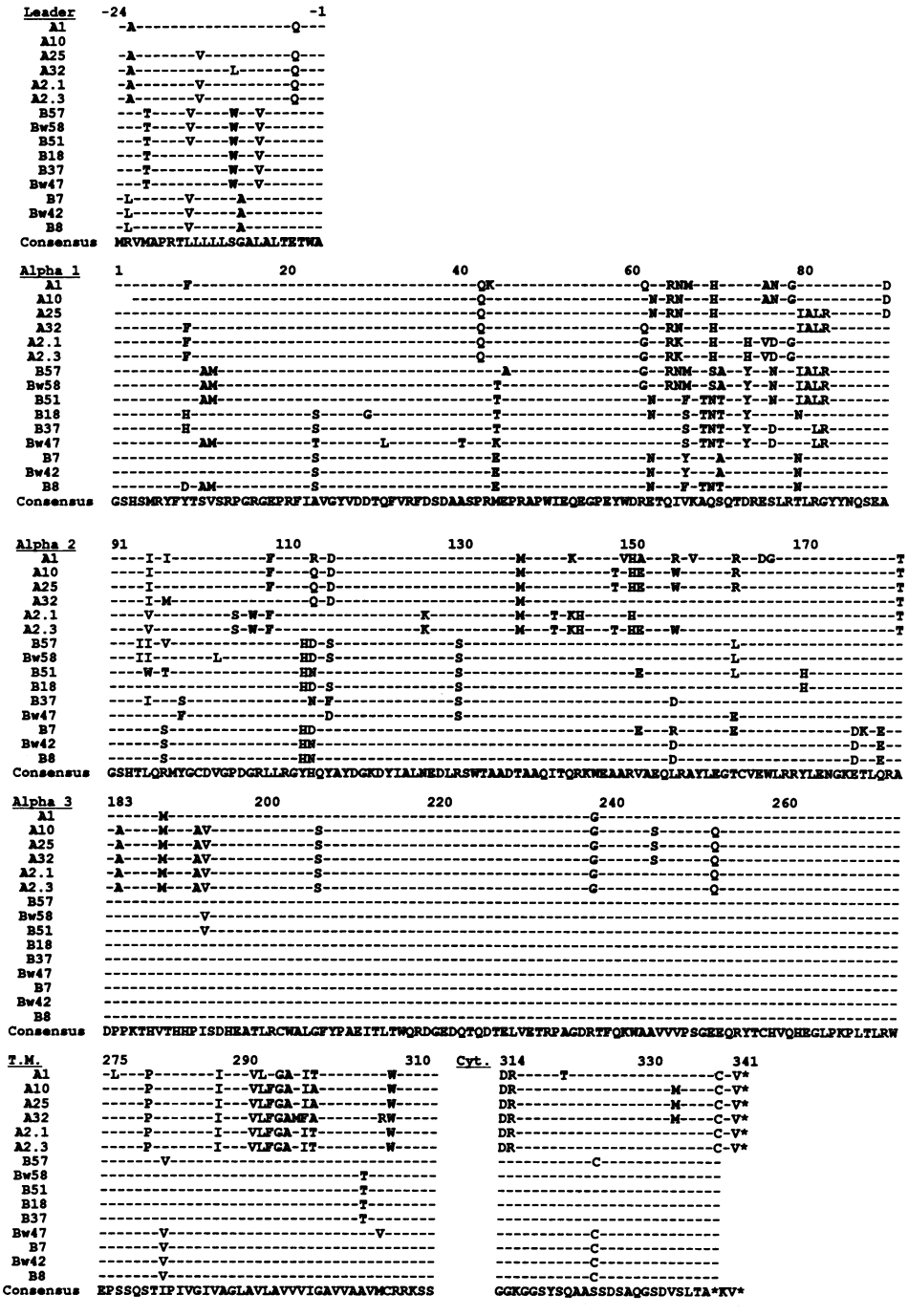


FIG. 4. Comparison of predicted protein sequences (in single-letter amino acid code) encoded by *HLA-A,B* alleles. The following sequences are from the cited references: HLA-A1, -A32, -B8, -B18, -Bw42 (5, 23); -A10 (19); -A2.1 (12); -A2.3 (24); -Bw58 (15); -B51 (20); and -Bw47 (25). The consensus sequence in α_1, α_2, and α_3 is identical to that given in Parham *et al.* (23) and is derived from 15 HLA-A, 20 HLA-B, and 4 HLA-C sequences. The consensus sequences for the leader, the cytoplasmic, and the transmembrane domains are derived from the sequences in the figure. Identities with the consensus sequence are given with a dash; asterisks indicate a termination codon.

10. Hanahan, D. (1985) in *DNA Cloning*, ed. Glover, D. M. (IRL, Oxford), Vol. 1, p. 109.
11. Ennis, P. D., Jackson, A. P. & Parham, P. (1988) *J. Immunol.* **141**, 642–651.
12. Koller, B. H. & Orr, H. T. (1985) *J. Immunol.* **134**, 2727–2733.
13. Sodoyer, R., Damotte, M., Delovitch, T. L., Trucy, J., Jordan, B. R. & Strachan, T. (1984) *EMBO J.* **3**, 879–885.
14. Kottmann, A. H., Seemann, G. H. A., Guessow, H. D. & Roos, M. H. (1986) *Immunogenetics* **23**, 396–400.
15. Ways, J. P., Coppin, H. L. & Parham, P. (1985) *J. Biol. Chem.* **260**, 11924–11933.
16. Güssow, D., Rein, R. S., Meijer, I., de Hoog, W., Seemann, G. H. A., Hochstenbach, F. M. & Ploegh, H. L. (1987) *Immunogenetics* **25**, 313–322.
17. Strachan, T., Sodoyer, R., Damotte, M. & Jordan, B. R. (1984) *EMBO J.* **3**, 887–894.
18. Bjorkman, P. J., Saper, M. A., Samraoui, B., Bennett, W. S., Strominger, J. L. & Wiley, D. C. (1987) *Nature (London)* **329**, 506–512.
19. Cianetti, L., Testa, U., Scotto, L., La Valle, R., Simeone, A., Boccoli, G., Giannella, G., Peschle, C. & Boncinelli, E. (1989) *Immunogenetics* **29**, 80–91.
20. Hayashi, H., Ennis, P. D., Ariga, H., Salter, R. D., Parham, P., Kano, K. & Takiguchi, M. (1989) *J. Immunol.* **142**, 306–311.
21. Sood, A. K., Pan, J., Biro, P. A., Pereira, D., Srivastava, R., Reddy, V. B., Duceman, B. W. & Weissman, S. M. (1985) *Immunogenetics* **22**, 101–121.
22. van Seventer, G. A., Spits, H., Yssel, H., Melief, C. J. M. & Ivanyi, P. (1988) *J. Immunol.* **141**, 417–422.
23. Parham, P., Lomen, C. E., Lawlor, D. A., Ways, J. P., Holmes, N., Coppin, H. L., Salter, R. D., Wan, A. M. & Ennis, P. D. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 4005–4009.
24. Holmes, N., Ennis, P., Wan, A. M., Denney, D. W. & Parham, P. (1987) *J. Immunol.* **139**, 936–941.
25. Zemmour, J., Ennis, P. D., Parham, P. & Dupont, B. (1988) *Immunogenetics* **27**, 281–287.
26. Salter, R. D., Norment, A. M., Chen, B. P., Clayberger, C., Krensky, A. M., Littman, D. R. & Parham, P. (1989) *Nature (London)* **338**, 345–347.
27. Orr, H. T. (1989) in *Immunobiology of HLA*, ed. Dupont, B. (Springer, New York), Vol. 2, pp. 33–40.