Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 11594–11597, December 1992
Immunology

Genomic polymorphism, recombination, and linkage disequilibrium in human major histocompatibility complex-encoded antigen-processing genes

P. M. VAN ENDERT*, M. T. LOPEZ*, S. D. PATEL*, J. J. MONACO†, AND H. O. MCDEVITT*

*Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA 94305-5402; and †Departments of Microbiology and Immunology, and Pathology, Medical College of Virginia, Virginia Commonwealth University, Richmond, VA 23298-0678

Contributed by H. O. McDevitt, August 6, 1992

ABSTRACT **Recently, two subunits of a large cytosolic protease and two putative peptide transporter proteins were found to be encoded by genes within the class II region of the major histocompatibility complex (MHC). These genes have been suggested to be involved in the processing of antigenic proteins for presentation by MHC class I molecules. Because of the high degree of polymorphism in MHC genes, and previous evidence for both functional and polypeptide sequence polymorphism in the proteins encoded by the antigen-processing genes, we tested DNA from 27 consanguineous human cell lines for genomic polymorphism by restriction fragment length polymorphism (RFLP) analysis. These studies demonstrate a strong linkage disequilibrium between *TAP1* and *LMP2* RFLPs. Moreover, RFLPs, as well as a polymorphic stop codon in the telomeric *TAP2* gene, appear to be in linkage disequilibrium with *HLA-DR* alleles and RFLPs in the *HLA-DO* gene. A high rate of recombination, however, seems to occur in the center of the complex, between the *TAP1* and *TAP2* genes.**

Screening for novel genes within the major histocompatibility complex (MHC) class II region recently led to the discovery of four genes that reside in a closely linked complex within a 50-kilobase (kb) region upstream of the *HLA-DO* gene or its rodent analogues. Among these, the *TAP1* and *TAP2* genes (1) [also termed *RING4* (2) or *PSF1* (3) and *RING11* (4) or *PSF2* (5) in humans] code for proteins belonging to a family of membrane transporters that possess an ATP binding cassette and six to eight transmembrane domains. Transfection experiments with a murine (6, 7) and a human (8) MHC class I-deficient cell line have demonstrated that both proteins are required for normal MHC class I assembly and expression. Immunoprecipitation experiments have shown that the product of the other two genes, the murine LMP2 and LMP7 [or RING12 (4) and RING10 (9) in humans] proteins belong to a large multicatalytic cytosolic protease complex, referred to as the proteasome (10, 11) or large multifunctional protease (LMP) (1).

Previous functional, biochemical, and molecular experiments had pointed to polymorphism in the putative processing genes; most significantly, a polymorphism in the rat *TAP2* gene involving 25 amino acid substitutions has been shown to alter significantly the spectrum of peptides eluted from MHC class I molecules (12), such that cytotoxic T cells from rat strains carrying a different *TAP2* allele mount a strong response to the "modified" class I molecule (13). Moreover, a polymorphism in the amino acid composition of the LMP2 and LMP7 proteins has been reported (14).

These findings prompted us to investigate the degree of polymorphism in the human *TAP* and *LMP* genes. We demonstrate a very limited genomic polymorphism in the MHC class II region antigen processing genes. Interestingly, however, linkage disequilibrium seems not only to exist within the complex of novel genes but also to extend from the processing genes to the *HLA-DR* region.

MATERIALS AND METHODS

Cell Lines. The homozygous typing lines, most of which were generated from lymphocytes of offspring of first-cousin marriages, have been described (15, 16). All lines have been deposited in the National Institute of General Medical Sciences Cell Repository, Institute for Medical Research, Camden, NJ. DP types had previously been obtained by cellular typing, while DR types were derived from serological, oligonucleotide, and restriction fragment length polymorphism (RFLP) typing.

cDNA Probes. The *LMP2* and *LMP7* probes were isolated by screening a λgt10 cDNA library from the homozygous human cell line WT51 (17) with a mouse *LMP2* cDNA (18) and *LMP7*-specific oligonucleotides (9), respectively. At a size of 0.8 kb (*LMP2*) and 1.1 kb (*LMP7*), both probes contain the entire open reading frame. The *TAP* and *DO* probes were isolated by screening of a cDNA library from the myelomonocytic human line U937 in pCDM8. The sequence of the 2.5-kb *TAP1* probe was identical to the published *RING4* sequence (2), beginning at nucleotide 315. The sequence of the 1.5-kb *DO* probe was identical to a published *DO* sequence (19). The sequence of the 1.5-kb *TAP2* probe was identical to the published *PSF2* sequence (5), beginning at nucleotide 1053; however, it encoded a threonine at codon 665 and termination at codon 687.

Southern Blots. Southern blots were prepared according to standard procedures and hybridized to cDNA probes that had been labeled to a specific activity of $>10^9$ cpm/μg with the random primer method. The restriction enzymes used were *Apa* I, *Apa*LI, *Ava* II, *Bam*HI, *Bgl* I, *Bgl* II, *Cfo* I, *Eco*NI, *Eco*RI, *Eco*RV, *Hha* I, *Hin*cII, *Hin*dIII, *Hin*fI, *Kpn* I, *Ksp* I, *Pst* I, *Pvu* II, *Sac* I, *Sau*3A1, *Sma* I, *Sty* I, and *Taq* I. All probes were hybridized to at least 19 of these digests.

Oligonucleotide Typing. To type the DNAs for the polymorphic coding sequence positions in *LMP2, TAP1*, and *TAP2*, 0.5–1.0 μg of genomic DNA was amplified in 30 cycles, using an annealing temperature of 42°C. For the *TAP1* codons 333 and 637 and *TAP2* codon 379, the oligonucleotides for PCR and hybridization were identical to the ones used by Colonna *et al.* (20). In the case of the *LMP2* codon 60, the PCR primers were two 18-mers corresponding to the *LMP2* coding sequence 50 base pairs (bp) upstream and downstream from codon 60. To type for the polymorphic codons 665 and 687 in *TAP2*, a 700-bp piece of genomic DNA including both codons was amplified. As typing oligonucleotides, 15-mers with the polymorphic nucleotide in the center

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; RFLP, restriction fragment length polymorphism.

were used. For hybridization, 10–20 μl of the PCR product was blotted onto nylon membranes and hybridized to the end-labeled sequence-specific oligonucleotides.

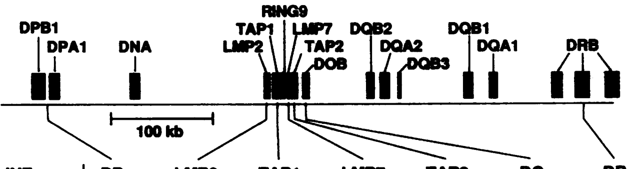
RESULTS AND DISCUSSION

To investigate the extent of genomic polymorphism and linkage disequilibrium in the *TAP* and *LMP* genes, we carried out RFLP analysis on DNA from 27 homozygous human lymphoblastoid cell lines that represent the Caucasian DR haplotypes DR1 through DR8. Because all but five of these lines are homozygous by consanguinity, the identification of individual alleles and linkage disequilibrium between different RFLPs is greatly facilitated (15, 16). DNA was digested with a panel of 19 restriction enzymes (Fig. 1) and sequentially hybridized to cDNA probes derived from the human *TAP* and *LMP* genes. We also included a probe for the *DO* gene (19), a HLA class II gene of unknown function that is located immediately telomeric of the *TAP2* gene.

In all putative processing genes, one of the 19 enzymes (in *TAP1, LMP7*, and *TAP2*) or two of them (in *LMP2*) detected polymorphic restriction sites subdividing the lines into major groups (Fig. 2). In the *LMP2* gene, diallelic RFLPs with an approximately equal prevalence of the two patterns were detected by a *Pst* I as well as an *Nco* I digest; with the exception of one line, allele typing of the lines by *Pst* I and *Nco* I RFLPs was identical, pointing to a strong association of these RFLPs. A similarly even distribution of a diallelic *Bgl* I polymorphism was detected in hybridizations to the *TAP1* probe, with an additional unique pattern in one line.

The most polymorphic pattern was observed in the *LMP7* gene, where a *Hha* I digest revealed six distinct fragment combinations. The observed combinations of two polymorphic *Eco*NI sites in the *TAP2* gene divided the lines into one frequent and two less frequent groups. In the *DO* gene, five enzymes detected RFLPs. Due to strong association of these polymorphisms, they could be clustered in three groups, which are represented by the *Sac* I digest shown in Fig. 1.

In addition to these widely distributed polymorphisms, RFLPs restricted to individual lines were found to a different extent for different genes. In a total of nine lines, unique RFLPs in the *LMP2* gene were detected by one (six lines), two (two lines) and six (one line) digests, respectively. Unique RFLPs in the *TAP1* gene were found in single digests of DNA from four lines. Only one line showed a unique RFLP in the *DO* gene, and no such patterns were detected in the *TAP2* and *LMP7* genes.



LINE	DP	LMP2	TAP1	LMP7	TAP2	DO	DR
JESTHOM	4	1	1	4	2	2	1
WT100	1	2	2	2	2	2	1
BVR*	2,4	1	1	2	2	2	1
IBW4	4	1	1	1	2	2	1
FJO*	n.d.	1,2	2	2	1	3	2
AKIBA	n.d.	1	1	3	1	1	2
WT18	3	1	1	1	1	1	2
WT87	3	2	2	2	1	3	2
IWB	2	2	2	2	1	3	2
MWF*	n.d.	1	1	2	2,3	2	3
HAR	4	2	2	n.d.	1	2	3
AVL	4	1	1	2	1	3	3
WT20	n.d.	1	1	4	1	3	3
WT49	4	2	2	3	3	2	3
TS10	n.d.	1	1	5	1	1	4
BIN40*	3,6	1	2	5	1	1	4
AL10	n.d.	1	1	5	1	1	4
WT51	2	2	2	5	1	1	4
JVM	2	1	1	4	1	3	5
FPF	2	2	2	3	1	1	5
ARENT	1	1	1	2	1	3	6
APD	4	2	2	6	3	2	6
HHK	4	2	2	2	3	2	6
MOU	2	1	1	2	1	1	7
IBW9	1	2	1	1	1	1	7
MADURA	4	2	3	2	1	2	8
BTB*	n.d.	2	2	2	1	1	8

FIG. 2. Location of *TAP* and *LMP* genes within the HLA class II region and distribution of major polymorphisms in homozygous lines. The genomic map is adapted from Trowsdale *et al.* (21). The lines were classified according to the RFLPs shown in Fig. 1. Stars indicate nonconsanguineous lines; n.d., not determined.

Thus, only very limited polymorphism is displayed by the human genomic regions encoding these antigen processing genes. A similarly limited degree of allelic variation has been

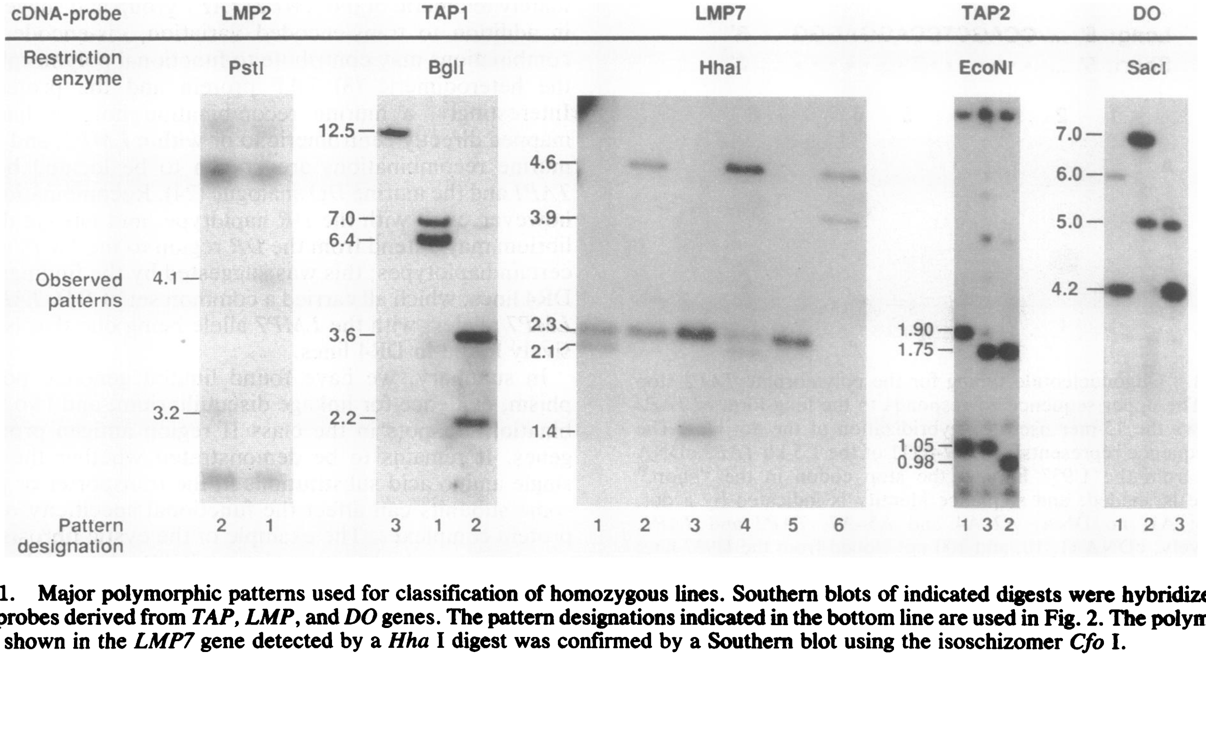


FIG. 1. Major polymorphic patterns used for classification of homozygous lines. Southern blots of indicated digests were hybridized with cDNA probes derived from *TAP, LMP*, and *DO* genes. The pattern designations indicated in the bottom line are used in Fig. 2. The polymorphic pattern shown in the *LMP7* gene detected by a *Hha* I digest was confirmed by a Southern blot using the isoschizomer *Cfo* I.

reported in a recent analysis of human *TAP1* and *TAP2* cDNAs, in which only two (*TAP1*) and three (*TAP2*) amino acid positions were found to be variable (20). An equivalent study of *LMP* cDNAs has not yet been published. A preliminary analysis, however, revealed that, besides a previously described arginine/histidine polymorphism (4), human LMP2 cDNAs display little variation (M.T.L. and S.D.P., unpublished observations).

Statistically, most of the polymorphic restriction sites reported here can be expected to reside within introns. We tested a possible association between RFLPs and previously published coding-sequence polymorphisms by typing of PCR-amplified genomic DNA with sequence-specific oligonucleotides. In 24 typed lines, arginine was found 17 times, and histidine 7 times, at amino acid 60 in the *LMP2* coding sequence. In the *TAP1* coding sequence, isoleucine was found more frequently than valine at position 333 (19 vs. 5 lines), while aspartic acid dominated over glycine at position 637 (22 vs. 3). Glycine-637 was associated with valine-333 in all cases. In the *TAP2* gene, the predominant valine at position 379 was replaced by isoleucine in 3 out of 24 typed lines. In the predicted *TAP2* ATP-binding cassette, 8 out of 27 lines possessed the combination of alanine-665 and glutamine-687 instead of the more frequent threonine-665 and a stop codon at 687. Isoleucine-379 was always associated with the latter set of codons. Thus, the frequencies and the associations of the polymorphic amino acids in the *TAP* coding sequences observed by us confirmed the findings by Colonna *et al.* (20).

When the results of RFLP and oligonucleotide typing were compared, a striking correlation was observed in the *TAP2* gene: as shown in Fig. 3, all 8 lines with one of the less frequent RFLP alleles, 2 and 3, carried the "long form" of *TAP2* (alanine-665 and glutamine-687). Moreover, the *Pst* I RFLP and the arginine/histidine polymorphism in codon 60 of the *LMP2* gene were in a strong association ($P = 0.012$). All 12 lines with the RFLP allele 1 carried an arginine, while the RFLP allele 2 was found associated with both histidine (6 lines) and arginine (5 lines). In the case of all other polymorphic codons (*TAP1* 333 and 637, *TAP2* 379), no such correlation was apparent. At least two explanations for this discrepancy are conceivable. First, analysis of a larger sample may reveal associations between the RFLPs and the coding-sequence polymorphisms. Second, recombination may occur within the processing genes.

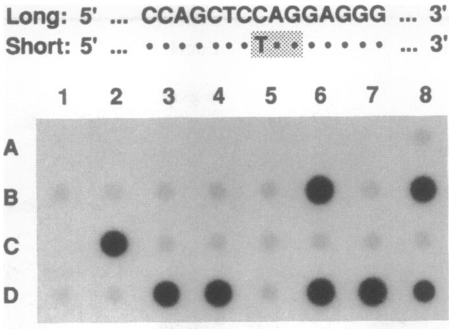

Long: 5' ... CCAGCTCCAGGAGGG ... 3'
Short: 5' ... •••••••T••••••• ... 3'
1
2
3
4
5
6
7
8
A
B
C
D

FIG. 3. Oligonucleotide typing for the polymorphic *TAP2* stop codon. The upper sequence corresponds to the long form of *TAP2* and shows the 15-mer used for hybridization of the dot blot. The lower sequence represents bp 1097–1111 of the 1.5-kb *TAP2* cDNA isolated from the U937 library; the stop codon in the "short" sequence is shaded, and sequence identity is indicated by a dot. Samples: A1, no DNA; A2–A4 and A5–A7, *TAP2* and *TAP1*, respectively, cDNA (1, 10, and 100 ng) cloned from the U937 line; A8–D8, PCR-amplified genomic DNA from homozygous lines, in the order TS10, JVM, Bin40, AL10, Akiba, WT18, MWF, Arent, Wt100, none (no DNA), BVR, FJO, AVL, FPF, WT87, Madura, WT51, IWB, WT20, IBw4, WT49, MOU, HHK, JESTHOM, HOM2.

Independent of their location in the genes, RFLPs can serve as markers for the study of linkage disequilibrium between two gene loci. To this end, we subjected the data shown in Fig. 2, whenever possible, to a statistical analysis for association between individual RFLP alleles in two genes or, alternatively, for a groupwise random association between alleles in two genes. The former procedure, which allows a formal analysis of linkage disequilibrium, could only be applied to a comparison of *TAP1* and *LMP2* RFLPs. In all other cases, the number of alleles and/or low frequency of individual alleles precluded application of this test; in these cases, the statistical evaluation was based on Fisher's exact test (22). A significant association in this test can be interpreted as a strong suggestion, but not proof, of linkage disequilibrium.

Significant linkage disequilibrium was found between the RFLP alleles in the *LMP2* and *TAP1* genes; a positive association was found between the alleles designated 1 ($P = 0.02$, χ^2 test) and 2 ($P < 0.02$). This linkage disequilibrium cannot be explained by the close genomic proximity of the two genes, since even closely linked loci are expected to recombine (given sufficient time) in the absence of linkage disequilibrium (23).

Using the less stringent evaluation by Fisher's exact test, RFLPs in *LMP2* and *TAP1* did not show an association with the alleles of the centromeric *DP* locus (*LMP2/DP*, $P = 1.0$) or with RFLPs in any of the telomeric *LMP7*, *TAP2*, or *DO* genes (*TAP1/LMP7*, $P = 0.22$; *TAP1/TAP2*, $P = 0.46$; *TAP1/DO*, $P = 0.24$). A significant association was found between RFLPs in the telomeric *TAP2* and *DO* genes ($P = 0.0034$). This association extended to include the *DR* alleles (*TAP2/DR*, $P = 0.0006$; *DO/DR*, $P = 0.0018$). All four DR1 lines, for example, carried an identical set of *DO* and *TAP2* alleles and possessed the "long form" of *TAP2*.

RFLPs in the central *LMP7* gene showed no significant association with RFLPs in any of the genes located upstream (*LMP7/TAP1*, $P = 0.22$) or downstream (*LMP7/TAP2*, $P = 0.36$; *LMP7/DO*, $P = 0.073$). Analysis of larger sample numbers, however, may reveal associations between these and other apparently randomly associated genes. In fact, an analysis of >150 DNAs from DR2 and DR3/4 individuals suggests that linkage disequilibrium may include the centromeric genes (R. Liblau and P.M.v.E., unpublished results). Nevertheless, these findings suggest that a high rate of recombination occurs both slightly centromeric and immediately telomeric of the *LMP2/TAP1* group and suggest that, in addition to trans-encoded variation, cis-encoded allele combinations may contribute to functional polymorphism in the heterodimeric (8) TAP protein and the proteasome. Interestingly, a murine recombination hotspot has been mapped directly centromeric to or within *LMP2*, and further murine recombinations are known to be located between *TAP1* and the murine *DO* analogue (24). Recombination may, however, vary with the *DR* haplotype, and linkage disequilibrium may extend from the *DR* region to the *LMP7* gene in certain haplotypes; this was suggested by the findings in the DR4 lines, which all carried a common set of *DO*, *TAP2*, and *LMP7* alleles, with the *LMP7* allele being one that is exclusively found in DR4 lines.

In summary, we have found limited genomic polymorphism, evidence for linkage disequilibrium, and two recombination hotspots in the class II region antigen processing genes. It remains to be demonstrated whether the scarce single amino acid substitutions in the transporter or proteasome subunits can affect the functional specificity of these protein complexes. The example of the cystic fibrosis transmembrane conductance regulator (CFTR) channel, however, shows that single substitutions can drastically alter the function of proteins belonging to the group of ABC transporters (25). Our observations of probable linkage disequilibrium between some processing genes, and between the processing genes and the *HLA-DR* region, have several implications. First, studies of polymorphism in the processing genes in MHC class II-associated diseases should include class II-matched controls. Second, the LMP2 and TAP1 proteins may functionally interact, resulting in selective pressure to maintain certain allele combinations. Finally, some of the processing genes (e.g., on the DR4 haplotype) may be selected as preferred functional partners for particular HLA protein alleles with which they are in linkage disequilibrium.

M.T.L. and S.D.P. contributed equally to this work. We thank Dan Denney for the WT51 library, Stephanie Land for the statistical analysis, and R. Liblau for helpful comments. P.M.v.E. was supported by a grant from the Deutsche Forschungsgemeinschaft, Bonn, F.R.G.; these studies were supported by the Human Genome Institute at the National Institutes of Health.

1. Bodmer, J. G., Marsh, S. G. E., Albert, E. D., Bodmer, W. F., Dupont, B., Erlich, H. A., Mach, B., Mayr, W. R., Parham, P., Sasazuki, T., Schreuder, G. M. T., Strominger, J. L., Svejgaard, A. & Terasaki, P. I. (1992) *Tissue Antigens* **39,** 161–173.
2. Trowsdale, J., Hanson, I., Mockridge, I., Beck, S., Townsend, A. & Kelly, A. (1990) *Nature (London)* **348,** 741–744.
3. Spies, T., Bresnahan, M., Bahram, S., Arnold, D., Blanck, G., Mellins, E., Pious, D. & DeMars, R. (1990) *Nature (London)* **348,** 744–747.
4. Kelly, A., Powis, S. H., Glynne, R., Radley, E., Beck, S. & Trowsdale, J. (1991) *Nature (London)* **353,** 667–668.
5. Bahram, S., Arnold, D., Bresnahan, M., Strominger, J. L. & Spies, T. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 10094–10098.
6. Spies, T. & DeMars, R. (1991) *Nature (London)* **351,** 323–324.
7. Attaya, M., Jameson, S., Martinez, C. K., Hermel, E., Aldrich, C., Forman, J., Fischer-Lindahl, K., Bevan, M. J. & Monaco, J. J. (1992) *Nature (London)* **355,** 647–649.
8. Powis, S. J., Townsend, A. R. M., Deverson, E. V., Bastin, J., Butcher, G. W. & Howard, J. C. (1992) *Nature (London)* **354,** 528–531.
9. Glynne, R., Powis, S. H., Beck, S., Kelly, A., Kerr, L.-A. & Trowsdale, J. (1991) *Nature (London)* **353,** 357–360.
10. Ortiz-Navarrete, V., Seelig, A., Gernold, M., Frentzel, S., Kloetzel, P. M. & Hämmerling, G. J. (1991) *Nature (London)* **353,** 662–664.
11. Brown, M. G., Driscoll, J. & Monaco, J. J. (1991) *Nature (London)* **353,** 355–357.
12. Powis, S. J., Deverson, E. V., Coadwell, W. J., Ciruela, A., Huskisson, N. S., Smith, H., Butcher, G. W. & Howard, J. C. (1992) *Nature (London)* **357,** 211–215.
13. Livingstone, A. M., Powis, S. J., Diamond, A. G., Butcher, G. W. & Howard, J. C. (1989) *J. Exp. Med.* **170,** 777–795.
14. Monaco, J. J. & McDevitt, H. O. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 3001–3005.
15. MacMurray, A., Bell, J. I., Denney, D., Watling, D., Foster, L. & McDevitt, H. O. (1987) *J. Immunol.* **139,** 574–586.
16. Bell, J. I., Denney, D., MacMurray, A., Foster, L., Watling, D. & McDevitt, H. O. (1987) *J. Immunol.* **139,** 562–573.
17. Bell, J. I., Denney, D., Foster, L., Belt, T., Todd, J. A. & McDevitt, H. O. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 6234–6238.
18. Martinez, C. K. & Monaco, J. J. (1991) *Nature (London)* **353,** 664–667.
19. Tonnelle, C., DeMars, R. & Long, E. O. (1985) *EMBO J.* **4,** 2839–2847.
20. Colonna, M., Bresnahan, M., Bahram, S., Strominger, J. L. & Spies, T. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 3932–3936.
21. Trowsdale, J., Ragoussis, J. & Campbell, R. D. (1991) *Immunol. Today* **12,** 443–446.
22. Mehta, C. R. & Patel, N. R. (1983) *J. Am. Stat. Assoc.* **78,** 427–434.
23. Bodmer, W. F. & Cavalli-Sforza, L. L. (1976) *Genetics, Evolution and Man* (Freeman, San Francisco).
24. Fischer-Lindahl, K. (1991) *Trends Genet.* **7,** 273–276.
25. Collins, F. S. (1992) *Science* **256,** 774–779.