Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 5860–5864, September 1985
Genetics

Intron sequences reveal evolutionary relationships among major histocompatibility complex class I genes

(gene conversion/coherent evolution/sequence homology/dendrogram/dispersed repeat elements)

HANS RONNE*†, EVA WIDMARK*, LARS RASK‡, AND PER A. PETERSON*

*Department of Cell Research, Uppsala University, Uppsala, Sweden; and ‡Swedish University of Agricultural Sciences, The Wallenberg Laboratory, Box 562, S-751 22 Uppsala, Sweden

Communicated by Peter Reichard, May 13, 1985

ABSTRACT **The multigene family of the class I histocompatibility antigens is unusual in that allelic and intergenic differences often are of equal magnitude. It has been suggested that this is due to gene conversion events, which would produce allelic variation but at the same time reduce intergenic differences. We compared the sequences of 11 class I genes in an attempt to elucidate the evolutionary history of this gene family. Our analysis shows that the intron sequences can be used to establish the order of divergence of various class I genes from each other. The results obtained agree with the order of divergence deduced from major insertion and deletion events. It appears that certain genes in the murine TL antigen-encoding region diverged very early from the *H-2* and *Qa-2,3* genes. The latter can be subgrouped as *H-2* and *Qa-2,3* genes by both sequence homology and insertion patterns. In contrast to the introns, exon sequences provide less information on evolutionary relationships. Thus, these analyses are consistent with the view that concerted evolution due to gene conversion occurs preferentially in exons.**

The basis of immunity is the ability to distinguish between foreign and self-determinants. In vertebrates this function is mediated by a group of related proteins that includes the immunoglobulins, the T-cell receptor(s), and the class I and class II major histocompatibility (MHC) antigens (1–4). Whereas the former two recognize foreign determinants, the MHC antigens are essential for self-recognition (restriction) in the immune response. Thus, foreign cell-surface antigens are recognized by the immune system only in the context of self MHC antigens. Class I restriction and class II restriction occur at different stages of the immune response. The recognition of a foreign antigen by helper T cells is class II-restricted, whereas the cytolytic attack of killer T cells is class I-restricted (5).

Class I molecules are membrane proteins with four extracellular domains. Three domains are located on a transmembrane chain, whereas the fourth domain is a separate protein, β_2-microglobulin (6). The transmembrane chains are highly polymorphic and are all encoded in the same chromosomal region, the MHC (2, 3). In the mouse, the MHC also contains a large number of class I genes that encode Qa and TL antigens (Fig. 1*a*). These are structurally related to the classical class I antigens (the H-2K, -D, and -L molecules) but are less polymorphic and are only expressed by differentiating lymphocytes (7). The human genome also contains many more class I genes than the number of human class I antigens identified so far (the HLA-A, -B, and -C molecules).

A conspicuous property of the H-2 and HLA antigenic systems is the fact that alleles are no more similar to each other than to nonallelic class I genes (8, 9). To account for this, it has been suggested that class I genes participate in gene conversion events involving short segments of DNA (9–11). Such conversions are likely to increase allelic polymorphisms at the expense of intergenic differences. The class I antigens would then be subject to concerted evolution, something which would frustrate attempts to analyze the evolutionary history of the gene family.

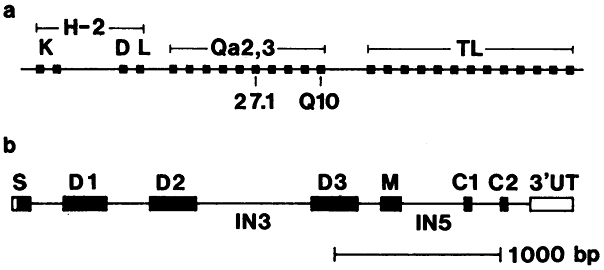

FIG. 1. Organization of MHC class I genes. (*a*) Schematic outline of the murine MHC on chromosome 17. The centromere is to the left. Only class I genes are shown in the figure (black boxes). The number of genes in each region varies between different inbred strains. (*b*) Exon/intron organization of a typical class I gene. Translated regions are black boxes, untranslated sequences are open boxes. The polypeptide domain encoded by each exon is abbreviated above the figure: S, signal peptide; D1–D3, extracellular domains; M, transmembrane peptide; C1–C2, cytoplasmic region. Introns (IN) 3 and 5 (discussed in the text) are marked below the figure.

There is some evidence that conversion events in class I genes occur preferentially in exons (9). In that case, comparisons of intron sequences could reveal evolutionary relationships that have been obscured in the coding regions. To investigate this possibility, we have aligned both coding and noncoding regions of those class I genes for which reasonably complete sequences are available. Our results confirm that exons provide little information on evolutionary relationships. However, the intron sequences seem to have evolved in a more conventional way, by duplications and mutational drift. Accordingly, they can be used to deduce the order of divergence of various class I genes from each other.

METHODS

Sequences Compared. The sequences of five murine class I genes have been published: the *H-2K^d* and *H-2K^b* alleles (9, 12), the *H-2L^d* gene (13, 14), the pseudogene *27.1*, and the *Q10* gene from the *Qa-2,3* region (15, 16). We have recently isolated and sequenced a class I gene from the A/J mouse. This gene, *T2A*, is a pseudogene from the TL-encoding region

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; bp, base pairs; kb, kilobase pairs.

†Present address: Department of Human Genetics and Development, College of Physicians and Surgeons, Columbia University, 701 W. 168th Street, New York, NY 10032.

(17), referred to here as *TL*. Another gene, *T1A*, adjacent to the *T2A* gene in the *TL* region has also been isolated, and the partial sequence of intron 5 has been determined (unpublished data). Four human class I genes have been sequenced: the *HLA-A3* and *HLA-CW3* genes (18, 19) and the pseudogenes *HLA-12.4* and *LN-11A* (20, 21). Finally, the sequence of a rabbit class I gene has been determined (22). These 11 sequences were used in the comparisons.

Alignment Methods. Low-stringency dot-matrix comparisons (six matches in eight positions) were used initially to identify regions of homology. Insertions and deletions were then identified by repeated alignments of all 11 sequences. Nucleotide substitution frequencies were calculated from pairwise comparisons of the aligned sequences, omitting those positions where either sequence had a deletion or an undetermined base.

Dendrogram. The data from the pairwise comparisons of intron sequences were corrected for multiple events and back mutations (23). The corrected values were used to compute a dendrogram (see Fig. 3). Distances within the dendrogram were optimized by using the least-squares method (24). Separate calculations were made for the *H-2/Qa-2,3* genes, the *HLA* genes, and interspecies comparisons. The error of each calculation is given in Fig. 3 as a variance (squared length of residual vector divided by the number of dimensions).

RESULTS

Alignment of Class I Genes. The sequences of the class I genes were aligned as detailed in the *Methods* section. Most differences between the aligned genes are either substitutions or small insertions or deletions. However, there are also some striking differences that distinguish subgroups among the genes. One such diagnostic feature is the length of intron 5 (Fig. 1*b*). This intron is about 400 base pairs (bp) in the *T1A* and *T2A* genes from the murine *TL* region (17). In contrast, the *H-2* and *Qa-2,3* genes have a much shorter intron. The longer intron is present in the human and rabbit genes, suggesting that a deletion occurred in a common ancestor of the *H-2* and *Qa-2,3* genes after the divergence of the *TL* genes. A similar difference is found in the second cytoplasmic exon, which has three extra codons at the 5′ end in the human, rabbit, and *T2A* genes. Again, the absence of these codons in the *H-2* and *Qa-2,3* genes suggests a deletion in a common ancestor.

Evolution of the Third Intron. A more complex picture is provided by the large third intron. In the human and rabbit genes, this intron is about 600 bp. The murine genes contain a core element of similar size that is homologous to the human and rabbit introns. However, a number of independent insertion events have added to the length of the murine introns (Fig. 2). Thus, the *T2A* gene has a 1200-bp insertion at the 5′ end of the core element, whereas the *H-2* and *Qa-2,3* genes have insertions at the 3′ end. The latter have diverged further by insertions that are specific to the *H-2* and *Qa-2,3* genes, respectively.

The *H-2*- and *Qa-2,3*-specific insertions are made up of dispersed repeat elements. Thus, both *H-2* and *Qa-2,3* genes have a B1 repeat (25) inserted in the middle of the intron. However, the points of insertion differ by 200 bp, suggesting two independent insertion events or a rearrangement in one group of genes. The *Qa-2,3* genes also have a tandem B2 repeat (26) at the 3′ end of the intron. An 8-bp target site duplication indicates that this tandem repeat was inserted as a single unit. There is some evidence that also the large 3′ and 5′ insertions in the murine genes were associated with dispersed repeats. Thus, within these insertions, DNA segments that are weakly homologous to dispersed repeat sequences are interspersed with A+T-rich stretches of DNA. Such stretches are often found at the end of dispersed repeat elements (27).

Sequence Comparisons. In pairwise comparisons, we counted the minimum number of nucleotide substitutions that separates any two sequences. Different mechanisms are probably responsible for substitutions and insertions or deletions. Since the latter two contribute significantly to divergence in introns but not in exons (28), care was taken to eliminate such events so as to count only true substitutions. Some typical values are shown in Table 1. There is a

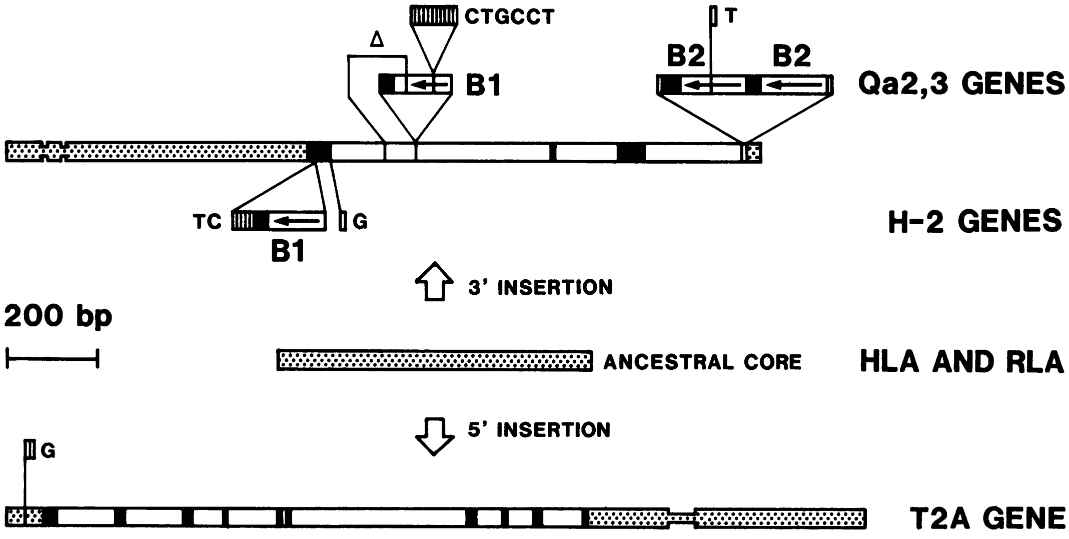

FIG. 2. Evolution of the third intron. The core element present in all class I genes is dotted. Black boxes are A- or T-rich stretches of DNA (6 of 8 matches). Inserted dispersed repeat elements are identified by an arrow and the name of the repeat family. Deletions in the core element are shown as constrictions. The Δ marks a deletion in the *Qa-2,3* genes that has removed part of the inserted B1 element. Hatched regions are simple DNA repeats, the extent of which varies from gene to gene. The simple repeat sequences are shown in the figure. The *Qa-2,3* genes are the *Q10* and *27.1* genes, whereas the *H-2* genes are the *H-2K*^{d} and *H-2L*^{d} genes. The sequences of the *Q10* and *H-2L*^{d} genes are not complete in this region. However, sufficient information is available to show that they have the same inserted repeat elements as the *27.1* and *H-2K*^{d} genes, respectively. In addition to the events shown, a large deletion has occurred in the *H-2L*^{d} intron. RLA, rabbit class I histocompatibility complex gene.

Table 1. Percent nucleotide substitutions between aligned class I genes

Genes compared	Exons 2–4	Introns 1–5
H-2K^d and *-K*^{b}	9.9	3.7
Q10 and *27.1*	10.9	6.1
H-2K^d and *-L*^{d}	9.8	7.8
H-2K^d and *27.1*	10.2	11.5
H-2L^d and *27.1*	10.7	13.2
HLA-A3 and *HLA-12.4*	8.4	9.2
HLA-A3 and *HLA-CW3*	9.7	15.2

Exons 2–4 encode the major extracellular part of the antigens. The number of nucleotide positions in the comparisons was 822 for the exons and 2417 for the introns.

pronounced difference between exons and introns. Thus, the *H-2/Qa-2,3* genes all differ from each other by approximately 10% in exons 2–4, a difference that equals that between nonallelic *HLA* genes. This is probably due to gene conversion events (9–11), which would tend to increase allelic polymorphism at the expense of intergenic *H2/Qa-2,3* differences.

In contrast, the introns show little evidence of concerted evolution. The *H-2K* alleles are more similar to each other than to the *H-2L* gene. The three *H-2* genes and the two *Qa-2,3* genes also form well-defined subgroups with respect to homology (Table 1). A similar difference is seen among the *HLA* genes, where the *HLA-CW3* gene has diverged from the other genes. These observations suggest that the evolution of the introns may be described by a conventional model in which duplicated or allelic genes diverge from each other by insertions, deletions, and mutational drift (29).

Dendrogram. If concerted evolution is an exon-specific process, then a comparison of intron sequences should reveal relationships that have been obscured in the coding regions. Accordingly, we used the intron data to compute a dendrogram (Fig. 3) in which the distance between any two sequences equals their separation in terms of PAM (percentage of accepted point mutations) units (23). Several conclusions can be drawn from this dendrogram. First, it confirms the observation that the rabbit gene is more closely related to the human genes than to the mouse genes (24, 30), suggesting a comparably more recent divergence of primates from lagomorphs.

Second, it shows that the *H-2* and *Qa-2,3* genes are all closely related to each other. In contrast, the *T2A* gene from the *TL* region differs considerably from the other mouse genes. In fact, this difference is as great as the difference between the human and rabbit genes. Thus, the divergence of the *T2A* gene from the *H-2/Qa-2,3* genes is probably at least as old as the divergence of primates from lagomorphs. This is a minimum estimate because it is possible that some exchanges (gene conversions) have occurred between the *H-2/Qa-2,3* and *TL* genes, resulting in a slower rate of intraspecies divergence. Limited sequence data on the *T1A* gene (intron 5) supports the notion that the *TL* genes differ from the *H-2/Qa-2,3* genes and also shows that the *T1A* and *T2A* genes have diverged considerably from each other.

A third conclusion is that the *H-2* genes and the *Qa-2,3* genes form two distinct subgroups. Among the *H-2* genes, the two *H-2K* alleles are most closely related to each other. In the *HLA* family the *HLA-CW3* gene stands well apart from the other genes. This divergence seems to be at least as old as the divergence of *H-2* from *Qa-2,3*. The *HLA-A3* gene differs as much from the two *HLA* pseudogenes as the *H-2L* gene differs from *H-2K*. The two pseudogenes, finally, are as similar to each other as are the two *H-2K* alleles.

DISCUSSION

Physical Extent of Concerted Evolution. The observation that class I antigens of different loci are no more divergent than allelic products of any given locus (8, 9) led to the suggestion and subsequent documentation that gene conversion-like events probably occur among the class I genes (9–11). As a corollary it is likely that class I genes are subject to concerted evolution, since frequent transfers of DNA segments between different loci would increase allelic poly-

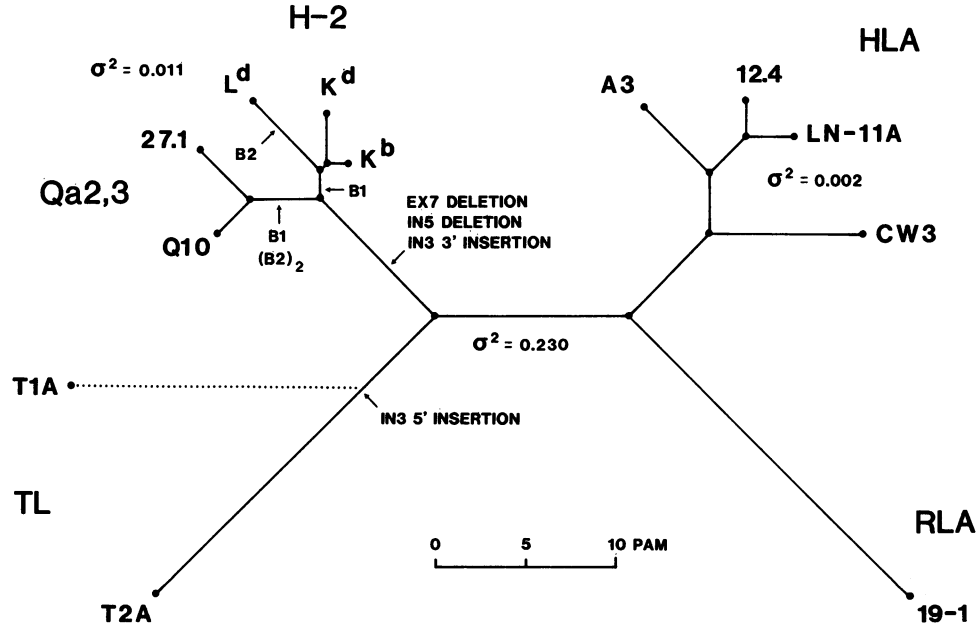

FIG. 3. Evolutionary dendrogram of MHC class I genes. The distance along the lines between any two genes is proportional to their divergence in PAM (percentage of accepted point mutations) units (23). The sequences of introns 1–5 were used for the comparisons. Introns 6 and 7 were omitted because of difficulties in making interspecies alignments. Also shown is the tentative position of the *T1A* gene (dotted line) based on the partial sequence of intron 5. The positions at which major insertions and deletions have occurred are marked by arrows. The variances shown are least squares. RLA, rabbit class I histocompatibility complex gene.

morphism but reduce interlocus differences. Since the putative conversion events are clustered in exons, mRNA intermediates may be involved in the process (9).

If this were the case, then concerted evolution due to gene conversion should be restricted to exons. Our results (Table 1; Fig. 3) are in agreement with the notion that concerted evolution is indeed most prominent in exons (9) and suggest that the noncoding sequences have evolved in a more conventional way. Even though this could be taken as support for an mRNA intermediate, another possible explanation is selection for antigenic polymorphism. Such selection would cause conversion events in exons 2–4 (which encode the major extracellular part of the molecule) to be preferentially retained in the population. New alleles isolated in the laboratory were also selected for antigenic differences. Thus, it is possible that conversions take place in both exons and introns, with only the former events being preserved.

Evolution of Class I Antigens. The fact that noncoding sequences evolve in a more conventional way makes it possible to trace the evolution of the class I genes. A dendrogram can be computed that gives the order of divergence of class I genes from each other (Fig. 3). It appears that the function and chromosomal organization of the class I genes reflect their evolutionary history. Thus, the *H-2K* and *H-2L* genes, which have a similar function, are more closely related to each other than to the *Q10* and *27.1* genes. Similarly, the two genes from the *Qa-2,3* region are more homlogous to each other than to the *H-2* genes.

More striking is the difference between the *T2A* gene and the *H-2/Qa-2,3* genes. The *TL* and *H-2/Qa-2,3* gene clusters have not been linked together (31) and, thus, probably are situated some distance apart on chromosome 17. This is in agreement with a comparably ancient divergence of the two regions from each other. Limited sequence data on the *T1A* gene shows that it shares at least one of the special features of the *T2A* gene. The sequencing of more genes from the *TL* cluster will reveal whether they are all more closely related to each other than to the *H-2/Qa-2,3* genes.

Insertions and Deletions. Strong independent support for the deduced evolutionary history is provided by the pattern of insertions and deletions. All such events can be fitted into the dendrogram in such a way that all genes beyond a given branch point share the event in question (Fig. 3). Thus, the genes in the *H-2/Qa-2,3* region have in common a deletion in intron 5, the absence of three codons in exon 7 and a large 3′ insertion in intron 3. Within this family, the *H-2K*^{d} and *H-2L*^{d} genes share a B1 insertion in intron 3, whereas the *Q10* and *27.1* genes have in common another B1 insertion and a tandem B2 element. Among the *H-2* genes, the *H-2L*^{d} and *H-2D*^{d} genes have a B2 insertion in the 3′ untranslated region that is not shared by the *H-2K* alleles (32). The *H-2L*^{d} gene also has a large deletion in intron 3 that is absent from the *H-2K* genes (14). Interestingly, insertions of dispersed repeat elements have occurred only in the murine genes. Possibly, these elements are more mobile in the mouse than in the human or rabbit genome. Alternatively, the insertion of one such element might facilitate further insertions.

Conclusions. The comparison of 11 class I gene sequences reveals that the intron sequences can be used to reconstruct the evolutionary history of this gene family. An evolutionary dendrogram computed from nucleotide substitutions in introns is in perfect agreement with the order of gene divergence determined independently from insertion and deletion events. Thus, both sequence homology and insertion patterns allow *H-2* and *Qa2,3* genes to be subgrouped among the class I genes. Moreover, at least some of the genes in the *TL* cluster reveal structural features that are only consistent with these genes having diverged very early from the *H-2/Qa-2,3* genes. While these evolutionary relationships are easily discerned by comparing intron sequences, analyses of exons are much less informative. This confirms and extends previous observations and may suggest that some mechanism for sequence transfer (such as gene conversion) contributes preferentially to the evolution of class I exons.

Note Added in Proof. After the completion of this work, the sequences of three more *H-2* genes have been published (33–35). As expected, their intron sequences are highly similar to the other *H-2* genes but less similar to the *27.1*, *Q10*, and *T2A* genes. A recent subgrouping of the murine class I genes based on hybridization data (36) is in agreement with the relationships outlined in Fig. 3.

We thank Dr. T. Kindt for making the sequence of the rabbit class I histocompatibility complex gene *19-1* available prior to its publication. This work was supported by grants from the Swedish Cancer Society.

1. Peterson, P. A., Rask, L., Sege, K., Klareskog, L., Anundi, H. & Östberg, L. (1975) *Proc. Natl. Acad. Sci. USA* **72**, 1612–1616.
2. Klein, J., Figueroa, F. & Nagy, Z. A. (1983) *Annu. Rev. Immunol.* **1**, 119–142.
3. Hood, L., Steinmetz, M. & Malissen, B. (1983) *Annu. Rev. Immunol.* **1**, 529–568.
4. Ploegh, H. L., Orr, H. T. & Strominger, J. L. (1981) *Cell* **24**, 287–299.
5. Doherty, P. C. & Zinkernagel, R. M. (1975) *J. Exp. Med.* **141**, 502–507.
6. Peterson, P. A., Rask, L. & Lindblom, J. B. (1974) *Proc. Natl. Acad. Sci. USA* **71**, 35–39.
7. Boyse, E. A. (1984) *Cell* **38**, 1–2.
8. Nathenson, S. G., Uehara, H., Ewenstein, B. M., Kindt, T. J. & Coligan, J. E. (1981) *Annu. Rev. Biochem.* **50**, 1025–1052.
9. Weiss, E. H., Golden, L., Zakut, R., Mellor, A., Fahrer, K., Kvist, S. & Flavell, R. A. (1983) *EMBO J.* **2**, 453–462.
10. Pease, L. R., Schulze, D. H., Pfaffenbach, G. M. & Nathenson, S. G. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 242–246.
11. Weiss, E. H., Mellor, A. L., Golden, L., Fahrner, K., Simpson, E., Hurst, J. & Flavell, R. A. (1983) *Nature (London)* **301**, 671–674.
12. Kvist, S., Roberts, L. & Dobberstein, B. (1983) *EMBO J.* **2**, 245–254.
13. Evans, G. A., Margulies, D. H., Camerini-Otero, R. D., Ozao, K. & Seidman, J. G. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 1994–1998.
14. Moore, K. W., Sher, B. T., Sun, Y. H., Eakle, K. A. & Hood, L. (1982) *Science* **215**, 679–682.
15. Steinmetz, M., Moore, K. W., Frelinger, J. G., Sher, B. T., Shen, F.-W., Boyse, E. A. & Hood, L. (1981) *Cell* **25**, 683–692.
16. Mellor, A. L., Weiss, E. H., Kress, M., Jay, G. & Flavell, R. A. (1984) *Cell* **36**, 139–144.
17. Hammerling, U., Ronne, H., Widmark, E., Servenius, B., Denaro, M., Rask, L. & Peterson, P. A. (1985) *EMBO J.* **4**, 1431–1434.
18. Sodoyer, R., Damotte, M., Delovitch, T. L., Trucy, J., Jordan, B. R. & Strachan, T. (1984) *EMBO J.* **3**, 879–885.
19. Strachan, T., Sodoyer, R., Damotte, M. & Jordan, B. R. (1984) *EMBO J.* **3**, 887–894.
20. Malissen, M., Malissen, B. & Jordan, B. R. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 893–897.
21. Biro, P. A., Pan, J., Sood, A. K., Kole, R., Reddy, V. B. & Weissman, S. M. (1983) *Cold Spring Harbor Symp. Quant. Biol.* **42**, 1082–1086.
22. Marche, P. N., Tykocinski, M. L., Max, E. E. & Kindt, T. J. (1985) *Immunogenetics (Heidelberg)* **21**, 71–82.
23. Dayhoff, M. O. (1978) *Atlas of Protein Sequence and Structure* (Georgetown Univ. Press., Washington, DC), Vol. 5, Suppl. 3, p. 375.
24. Kreider, D. L., Kuller, R. G., Ostberg, D. R. & Perkins, F. W. (1966) *An Introduction to Linear Analysis* (Addison-Wesley, Reading, MA), pp. 290–295.
25. Krayev, A. S., Kramerov, D. A., Skryabin, K. G., Ryskov, A. P., Bayev, A. A. & Georgiev, G. P. (1980) *Nucleic Acids Res.* **8**, 1201–1215.
26. Krayev, A. S., Markusheva, T. V., Kramerov, D. A.,

Ryskov, A. P., Skryabin, K. G., Bayev, A. A. & Georgiev, G. P. (1982) *Nucleic Acids Res.* **10,** 7461–7475.

27. Jelinek, W. R. & Schmid, C. W. (1982) *Annu. Rev. Biochem.* **51,** 813–844.

28. Efstratiadis, A., Posakony, J. W., Maniatis, T., Lawn, R. M., O'Connel, C., Spritz, R. A., DeRiel, J. K., Forget, B. G., Weissman, S. M., Slightom, J. L., Blechl, A. E., Smithies, O., Baralle, F. E., Shoulders, C. C. & Proudfoot, N. J. (1980) *Cell* **21,** 653–668.

29. Wilson, A. C., Carlson, S. S. & White, T. J. (1977) *Annu. Rev. Biochem.* **46,** 573–639.

30. Tykocinski, M. L., Marche, P. N., Max, E. E. & Kindt, T. J. (1984) *J. Immunol.* **133,** 2261–2269.

31. Weiss, E. H., Golden, L., Fahrner, K., Mellor, A. L., Devlin, J. J., Bullman, H., Tiddens, H., Bud, H. & Flavell, R. A. (1984) *Nature (London)* **310,** 650–655.

32. Kress, M., Barra, Y., Seidman, J. G., Khoury, G. & Jay, G. (1984) *Science* **226,** 974–977.

33. Arnold, B., Burgert, H.-G., Archibald, A. L. & Kvist, S. (1984) *Nucleic Acids Res.* **12,** 9473–9487.

34. Sher, B. T., Nairn, R., Coligan, J. E. & Hood, L. E. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 1175–1179.

35. Morita, T., Delarbre, C., Kress, M., Kourilsky, P. & Gachelin, G. (1985) *Immunogenetics (Heidelberg)* **21,** 367–383.

36. Rogers, J. H. (1985) *Immunogenetics (Heidelberg)* **21,** 343–353.