Proc. Natl. Acad. Sci. USA
Vol. 86, pp. 9986–9990, December 1989
Genetics

Ancient roots for polymorphism at the HLA-DQα locus in primates

(polymerase chain reaction/evolution/selection)

ULF B. GYLLENSTEN AND HENRY A. ERLICH

Department of Human Genetics, Cetus Corporation, 1400 Fifty-Third Street, Emeryville, CA 94608

Communicated by L. L. Cavalli-Sforza, August 25, 1989

ABSTRACT **The genes encoding the human histocompatibility antigens (HLA) exhibit a remarkable degree of polymorphism as revealed by immunologic and molecular analyses. This extensive sequence polymorphism either may have been generated during the lifetime of the human species or could have arisen before speciation and been maintained in the contemporary human population by selection or, possibly, by genetic drift. These two hypotheses were examined using the polymerase chain reaction method to amplify polymorphic sequences from the DQα locus, as well as the DXα locus, an homologous but nonexpressed locus, in a series of primates that diverged at known times. In general, the amino acid sequence of a specific human DQα allelic type is more closely related to its chimpanzee or gorilla counterpart than to other human DQα alleles. Phylogenetic analysis of the silent nucleotide position changes shows that the similarity of allelic types between species is due to common ancestry rather than convergent evolution. Thus, most of the polymorphism at the DQα locus in the human species was already present at least 5 million years ago in the ancestral species that gave rise to the chimpanzee, gorilla, and human lineages. However, one of the DQα alleles may have arisen after speciation by recombination between two ancestral alleles.**

The human histocompatibility class II genes encode three cell-surface antigens, designated HLA-DR, HLA-DQ, and HLA-DP. Each antigen consists of an α and a β chain. The allelic sequence diversity resides mainly in the second exon, which encodes the amino-terminal domain (1–4). Eight allelic variants have been found at the HLA-DQα locus (also known as *DQA1*) (3, 5, 6), and these alleles have been classified into four major types, designated A1–A4. An homologous and nonexpressed locus denoted DXα (also known as *DQA2*) has also been found (7), and appears to be monomorphic (3). With polymerase chain reaction primers designed for the polymorphic second exon of the HLA-DQα locus (5, 8) we have amplified the homologous DNA segment from nine chimpanzees (*Pan troglodytes*), four gorillas (*Gorilla gorilla*), and two individuals each of baboon (*Papio leucophaeus*), rhesus (*Macaca mulatta*), langur (*Presbytis entellus*), capuchin monkey (*Cebus capucinus*), and marmoset (*Callithrix* spp.). These species are a set of primates whose divergence times from the human lineage, estimated from fossil data as well as from molecular comparisons, rise gradually from the ≈5 million years (Myr) ago for *Pan* and *Gorilla* to nearly 40 Myr ago for the New World species (*Cebus* and *Callithrix*) (9–15). This set of species allowed us to test hypotheses concerning the age of the polymorphism at the DQα locus, as well as to study the mechanism primarily responsible for generating variation at these loci.

MATERIALS AND METHODS

One microgram of genomic DNA was subjected to 30 cycles of polymerase chain reaction (16–18) by using the primers and amplification conditions previously described for human DNA samples (5, 8). By contrast, these primers cannot amplify the DQα locus from more distantly related mammals such as sheep (*Ovis ovis*) and horse (*Equus equus*) or fish (brown trout, *Salmo trutta*) (unpublished work). The sequence variation in the 242-nucleotide-pair amplified fragment was first assayed by denaturing gradient gel electrophoresis (19) to distinguish between homozygous and heterozygous individuals. Homozygous individuals could then be selected for direct sequence analysis by using the protocol described for generating single-stranded DNA (5). Amplified DNA from heterozygous individuals was cloned into M13 vectors and sequenced or the individual alleles were purified from denaturing gradient gels for subsequent direct sequencing. Amplified DNA fragments to be cloned were digested with *Bam*HI and *Pst* I and ligated into M13mp18, and the recombinant phages were identified by plaque hybridization.

The sequences were derived from 30 humans, 9 chimpanzees, 4 gorillas, and 2 individuals of each of the other primate species. In our previous work on HLA class II sequence polymorphism, we used the locus nomenclature DQα and DXα and designated the alleles at the DQα locus A1–A4, with the subtypes of A1 denoted A1.1, A1.2, and A1.3. Recently a new system of nomenclature for the class II loci has been introduced in which DQα is designated *DQA1* and DXα is designated *DQA2* (20). To avoid confusion between the locus and allele description we have in this paper retained our previous nomenclature.

Maximum parsimony trees (21, 22) were constructed by using the phylogenetically informative positions in the amino acid sequence or the third positions of the codons in the nucleotide sequences and rooted at the midpoint of the branch separating the most distantly related sequences. All character changes, including deletions, were given equal weight. Branch lengths are not proportional to the number of changes. As a measure of the homoplasy of the tree, caused by convergent evolution or parallel mutation, we calculated the weighted sum of the minimum number of changes for each character divided by the weighted sum of the observed number of changes for each character. This index varies from 0–1, with 1 indicating absence of homoplasy.

RESULTS

Five different DNA sequences were found in the chimpanzee and the gorilla and two in each of the other species; their corresponding amino acid sequences appear in Fig. 1. For chimpanzee and gorilla, the sequences showed extensive similarity to the human DQα allelic types A1, A3, and A4 as well as to DXα. No sequence was found that was more closely related to the relatively rare human *DQA2* allele than to the other alleles (6). In humans, the *DQA2* allele is found only on DR7 haplotypes, consistent with the notion of its recent creation; it is the only DQα allele uniquely associated with a specific DR haplotype. Based on our phylogenetic analysis (Fig. 2), the *DQA2* allele may have been derived

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: Myr, million years.

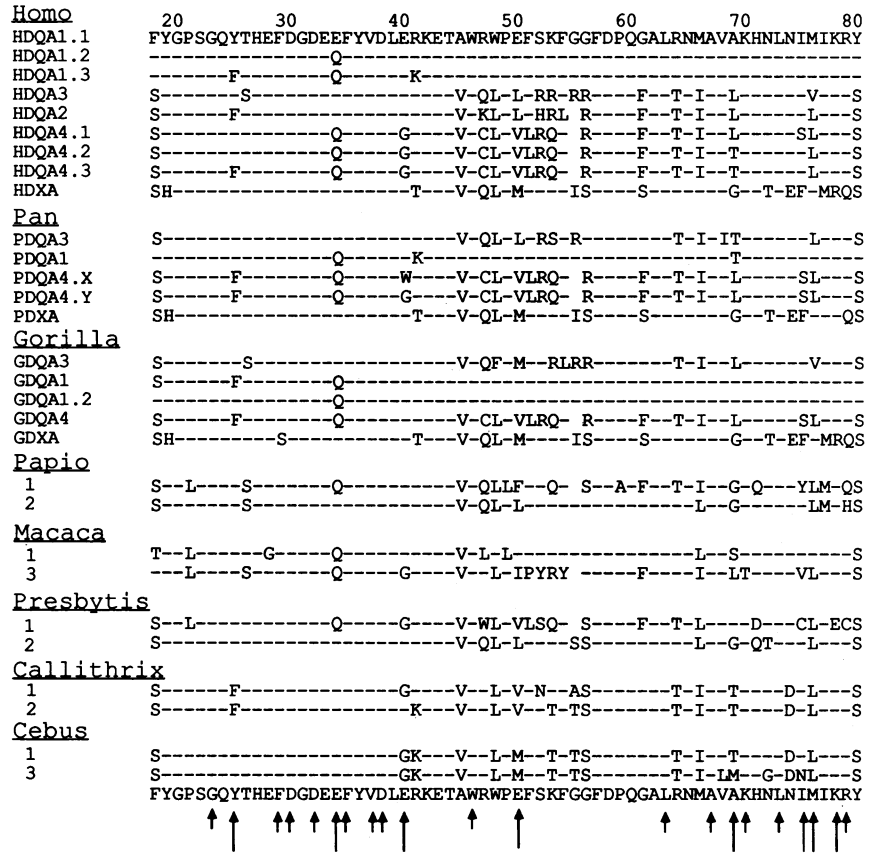

FIG. 1. Alignment of amino acid sequences of the second exon of the DQα locus and DXα from human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*), baboon (*Papio leucophaeus*), rhesus (*Macaca mulatta*), langur (*Presbytis entellus*), capuchin monkey (*Cebus capucinus*), and marmoset (*Callithrix* spp.). Long arrows at bottom indicate polymorphic residues, and short arrows indicate conserved residues in the alignment of class I and class II sequences (23). In a recent model of class II structure, variation in the right portion of the amplified DQα segment can be tentatively assigned to one of the α helices, whereas part of the left portion is predicted to form a β-pleated sheet (23).

from a *DQA3* progenitor sequence after speciation (see *Discussion*). The only amino acid residues unique to humans are the lysine at position 47 and the histidine at position 52 of the *DQA2* allele. In addition to the primates studied here by sequence analysis, the analysis of DQα polymorphism by oligonucleotide probe hybridization of an additional 19 individuals (chimpanzee, gorilla, and orangutan) revealed the presence of *DQA1*, *DQA3*, and *DQA4* but no *DQA2* alleles (A. Bowcock and J. Kurtz, personal communication).

The sequences in Fig. 1 reveal a number of cases where a given human allelic type is more closely related to its chimpanzee or gorilla counterpart than to other human alleles; the most extreme being the *DQA1.2* allele from the gorilla the amino acid sequence of which is identical to that of the human *DQA1.2* allele (Fig. 1). To determine whether the amino acid sequences of the alleles are more similar *within* species than *between* species, indicating recent divergence of the human alleles, or vice versa, we used a phylogenetic analysis based on maximum parsimony. The most parsimonious tree for the amino acid sequences from *Homo*, *Pan*, and *Gorilla* requires 62 changes and shows the extensive similarity of alleles from different species, presumably because allelic sequence divergence predates speciation (Fig. 2*a*). A tree where the sequences cluster according to their species origin, representing the hypothesis that the variation was generated after the split of the species, requires a total of 128 character changes (data not shown).

The observed similarity of alleles from different species may reflect either common ancestry or convergent evolution (24). However, the most parsimonious tree for the hominoids based on silent changes (Fig. 2*b*) is similar to that based on amino acid sequences in that it also links sequences from different species (Fig. 2*a*). The most parsimonious tree based on silent changes requires 36 changes, compared with 65 changes when the sequences form three species clusters, indicating that the similarity of alleles between species is due to a common ancestry predating the split of species. The clustering of DQα alleles in the silent tree (Fig. 2*b*) is also similar to the relationships defined by the supertypic serologic DR specificities in that, in humans, the closely related *DQA2* and *DQA3* alleles are linked to the DRw53 type (DR4, 7, and 9), the *DQA4* alleles to the DRw52 (DR3, 5, and 8), and the *DQA1* alleles are found on DR1 and DR2 haplotypes.

Tentative estimates of the age of the DQα allelic types were calculated from both the silent nucleotide substitutions that have accumulated *between* allelic types (e.g., between *DQA4* and *DQA1*) and the divergence time of the species estimated from the fossil data. The number of silent changes indicate that the allelic types split ≈20–40 Myr ago (for example, A1–A4, 26–40 Myr ago; A1–A3, 23–31 Myr ago), assuming a substitution rate of 0.12–0.16% per nucleotide pair per million years (12, 15). The presence of *DQA1*-, *DQA3*-, and *DQA4*-like sequences in all the hominoids shows that these allelic types are at least 5 Myr old (14). Thus, both the molecular estimates and those derived by methods that are independent of the molecular clock consistently show that the allelic types in the contemporary human population predate the divergence of *Homo*, *Pan*, and *Gorilla*. The phylogenetic assignment of some of the other primate sequences (e.g., *Presbytis* sequence 1, A4; or *Macaca* sequence 1, A1) suggests that these allelic types may, in fact, be considerably older.

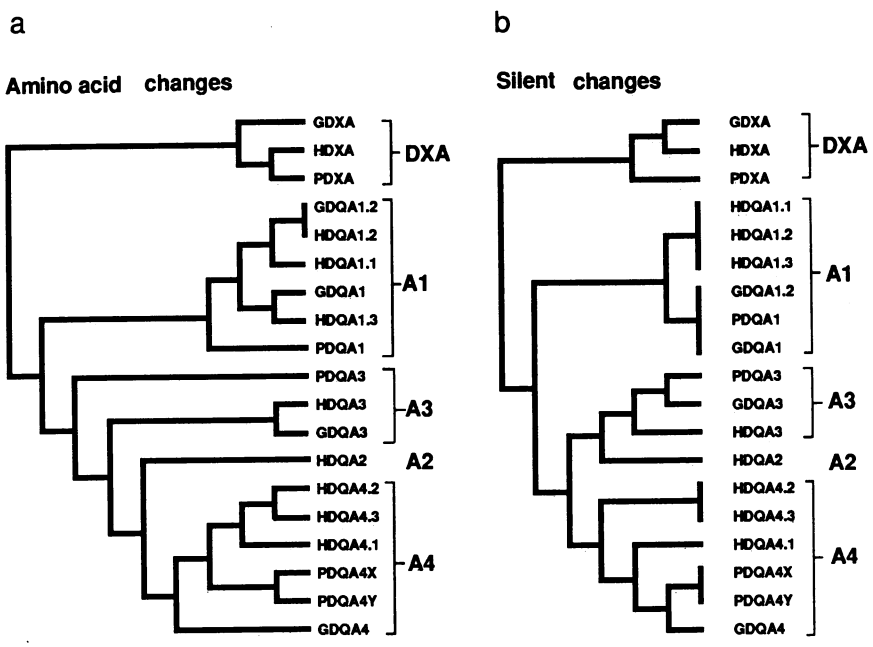

FIG. 2. Parsimony trees for the DQα and DXα (DXA) sequences from human (*Homo*), chimpanzee (*Pan*), and gorilla (*Gorilla*). (*a*) The most parsimonious tree for the sequences from *Homo, Pan*, and *Gorilla* based on 30 phylogenetically informative amino acid positions. This tree requires 62 character changes (homoplasy index, 0.77). A tree based on the same set of data where the sequences form three groups reflecting their species origin requires 128 character changes (homoplasy index, 0.38). (*b*) Tree based on phylogenetically informative variation at the third nucleotide position of 21 codons. This tree requires 36 changes (homoplasy index, 0.67) compared with 65 changes (homoplasy index, 0.37) when the sequences form three species clusters. The topology given suggests that the gene duplication originating the DXα locus predates the diversification of the DQα alleles (7). However, equally parsimonious topologies are possible that join the DX and the *DQA1* lineages. This fact may indicate either a common origin for the DX/*DQA1* sequences or the possibility of sequence exchange between the two loci. Bootstrapping of the amino acid sequence data indicates that the topology most consistent between subsets of the data positions the DX outside the DQ alleles.

These age estimates allow us to reject the hypothesis that the alleles are neutral. The time at which two neutral alleles diverged from an ancestral sequence is $\approx 4N_e$ generations (25), where N_e is the long-term effective population size for the species. For DQα, $4N_e$ = 5–20 Myr, which, with a generation time of 5 yr, equals 1–4 million generations. The effective population size required to retain the DQα alleles if they were completely neutral is, therefore, 250,000–1 million, a number unrealistic for most of these primate species.

DISCUSSION

It has been proposed that the extensive polymorphism observed at HLA loci has been generated, in part, by mechanisms such as gene conversion or segmental transfer (26). By combining fragments from different preexisting alleles, the overall similarity of alleles *within* species would increase and, if these exchanges were frequent enough, they would eventually obliterate the similarity of alleles between species. If the polymorphism was generated mainly by point mutations and interallelic recombination were rare, allelic types should be more similar *between* species. The hominoid DQα sequences show no indication of frequent segmental transfer or gene conversion over a time span of 5 Myr. However, the sequences *Papio* 2 and *Presbytis* 2 could have been generated by combining the first half of an A3-like sequence with the last half of an A1-like sequence (Fig. 1). These sequences therefore do not group with any of the major types and have not been included in the tree based on all phylogenetically informative nucleotide positions (Fig. 3). Similarly, the *DQA2* allele, unique to humans, may have arisen by intra-exon recombination between the A3 and A4 allelic types. Due to the limited number of alleles analyzed in these species it is hard to estimate the relative importance of gene conversion and point mutations. However, some of the other Old World monkey sequences show similarity to a specific hominoid allele (e.g., *Presbytis* sequence 1 to type A4). By contrast, the New World monkey sequences show little similarity with any particular hominoid sequence, and they form a separate branch in the phylogenetic tree (Fig. 3).

The effect of selection was inferred from the ratio of amino acid replacement to silent nucleotide substitutions in comparing sequences of the same allelic type between species.

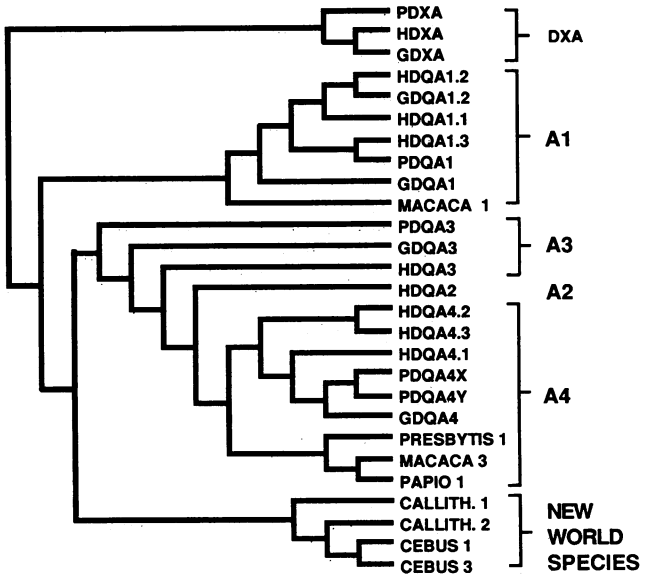

FIG. 3. Parsimony tree for the primate DQα and DXα sequences, based on 66 phylogenetically informative nucleotide positions. This tree requires 176 character changes (homoplasy index, 0.55). All sequences except *Papio* 2 and *Presbytis* 2 (Fig. 1) have been included in the tree. Alternative trees for positioning the Old World monkey sequences require an additional 6–13 character changes (e.g., *Macaca* 1 with *DQA2*, *-3*, and *-4*: 6 changes; *Macaca 3 with DQA1*: 13 changes; *Presbytis* 1 with *DQA1*: 12 changes; *Papio* 1 with *DQA1*: 8 changes).

The ratio of replacement to silent changes for the hominoid A1 allelic type is close to the ratio observed for the nonexpressed DXα locus and conforms to the 3:1 ratio found for nuclear genes under weak functional constraints (Table 1). Thus, the DQα alleles are evolving at a comparable rate (e.g., *DQA1*) or more slowly (e.g., *DQA4*) than the DXα gene, suggesting that they have been conserved by selection and not subjected to positive selection for variation, which should increase the ratio relative to a nonexpressed gene. These results, as well as those of other recent studies on major histocompatibility complex polymorphism (27–32), are consistent with a model of ancient allelic diversity and in contrast to the expectation for recently generated polymorphisms and strong selection maintaining newly arisen mutations.

Some hominoid alleles have accumulated fewer replacement changes than that expected from the number of silent differences (Table 1). For example, *DQA4*, next to *DQA3*, has accumulated the largest number of silent changes; yet, its amino acid sequence is the most conserved of all allelic types examined. The selection conserving the amino acid sequence of the α chain encoded by this allele could possibly be imposed by the necessity for the A4 α chain to pair with several different β chains. In humans, nonrandom haplotypic association of DQα and DQβ alleles has been seen (Table 2). Experiments with transfected cells also suggest that certain human DQα chains are unable to pair with some DQβ chains (37). In the mouse, the preferential association of specific combinations of A α chains and β chains (murine homologues to the DQ heterodimer) has been demonstrated (33) and mapped to polymorphic residues in the amino-terminal domain of both the α and β chains (34).

This putative selective constraint on A4 variability may extend to the other DQα alleles (Table 2). The α chains encoded by the *DQA4* and *DQA3* alleles (with a low ratio) pair with a diverse group of β chains, whereas the α chain encoded by the A1 allele (high ratio) pairs with a more restricted set of β chains. Thus, there appears to be an inverse relationship between the variability of the α chain between species and the diversity of the associated β chains, suggesting that selection is acting to conserve the functional interaction of the α and β chains. Consistent with this hypothesis is the observation, in humans, that the DRα (one allele) and DPα (two alleles) chains, both of which pair with many different β chains ($\approx$25–30), show very restricted polymorphism. Also, the ratio of replacement to silent changes for the DRα (or DRA) locus of humans and chimpanzee is low (35). Although our hypothesis to account for the observed differences in DQα evolutionary rates is based on the specific combinations of α- and β-chain alleles found on human haplotypes, the analysis of DQβ chain variation in the primates supports the conservation of certain haplotypic combinations of α and β chains (unpublished work).

In summary, most allelic diversity at the DQα locus, in contrast to the evolution of the DQβ and DRβ loci (unpublished work), in hominoids appears to have preceded speciation, and the rate of accumulation of amino acid substitutions is comparable to that of other nuclear genes. The maintenance of polymorphism at the DQα locus may be due predominantly to either overdominant (36) or frequency-dependent selection for ancient alleles rather than the operation of a mechanism for generating new variants by interallelic recombination. Further, different alleles at the DQα locus appear to be acquiring replacement substitutions at different rates. To account for this, we have proposed that the diversity of β chains with which an α chain pairs may constrain the α chain from evolving freely.

Table 1. Average number of amino acid and silent nucleotide-pair differences between DQα and DXα sequences

	Homo vs. *Pan*	*Homo* vs. *Gorilla*	*Pan* vs. *Gorilla*	Average (*Pan*, *Homo*, and *Gorilla*)	Average (*Homo* vs. OW monkeys)
DXα*					
aa	2	2	4	2.7	
snp	1	1	1	1	
DQA1					
aa	2.3	1.7	3	2.3	10†
snp	1	1	0	0.7	3
DQA3					
aa	7	5	10	7.3	
snp	6	3	6	5	
DQA4					
aa	2.5	3	1	2.1	12.3‡
snp	3.5	1.3	3	2.4	8

The numbers are based on averages of all possible pairwise combinations within allelic type. It is difficult to assess the statistical significance of differences in the ratio of amino acid (aa) to silent nucleotide-pair (snp) changes. A nonparametric test of the ratio of replacement to silent changes in all pairwise combinations of *DQA1* alleles and *DQA4* alleles was done. In this test the ratios were ranked according to size, and the number of runs of *DQA1* and *DQA4* comparisons was computed. The probability of equal ratios for the A1 and A4 type was $P < 0.025$. OW, Old World.

*DXα sequences were found only in *Homo*, *Pan*, and *Gorilla*.

†Comparison between *DQA1.1* and *Macaca* sequence 1.

‡Average of comparisons between *HDQA4.1* vs. *Presbytis* sequence 1, *Papio* sequence 1, and *Macaca* sequence 3.

Table 2. Haplotype association of alleles at the DQα and DQβ locus in human

α chain	β chain
DQA1.1	*DQB1.1* (Dw1), *−1.3* (Dw9)
DQA1.2	*DQB1.2* (Dw21), *−1.3* (Dw9), *−1.5* (Dw2), *−1.7* (Dw19)
DQA1.3	*DQB1.4* (Dw12), *−1.6* (Dw18)
DQA2	*DQB2*, *DQB3.3*
DQA3	*DQB3.1*, *−3.2*, *−3.3*, *DQB2* (Black DR7), *DQB4* (Japanese DR4)
DQA4	*DQB2*, *DQB3.1*, *DQB4*

We thank Allan C. Wilson and Jeff Hall for providing DNA samples and Corey Levenson, Dragan Spasic, and Lauri Goda for synthesis of oligonucleotides. We are grateful to Allan C. Wilson, Norm Arnheim, Ann Begovich, Russ Higuchi, John Sninsky, Mark Stoneking, Tom White, and Liz Zimmer for helpful discussions. U.B.G. was supported by a Postdoctoral Fellowship from the Knut and Alice Wallenberg Foundation (Sweden).

1. Trowsdale, J., Young, J. A. T., Kelly, A. P., Austin, P. J., Carson, S., Meunier, H., So, A., Erlich, H. A., Spielman, R., Bodmer, J. & Bodmer, W. F. (1985) *Immunol. Rev.* **85**, 5–43.
2. Giles, R. C. & Capra, J. D. (1985) *Adv. Immunol.* **37**, 1–71.
3. Horn, G., Bugawan, T. L., Long, C. M. & Erlich, H. A. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 6012–6016.
4. Bugawan, T. L., Horn, G. T., Long, C. M., Mickelson, E., Hansen, J. A., Ferrara, G. B., Angelini, G. & Erlich, H. A. (1988) *J. Immunol.* **141**, 4024–4030.
5. Gyllensten, U. & Erlich, H. A. (1988) *Proc. Natl. Acad. Sci. USA* **85**, 7652–7656.
6. Higuchi, R., von Beroldingen, C. H., Sensabaugh, G. F. & Erlich, H. A. (1988) *Nature (London)* **332**, 543–546.
7. Auffray, C., Lillie, J. W., Arnot, D., Grossberger, D., Kappes, D. & Strominger, J. L. (1984) *Nature (London)* **308**, 327–333.
8. Scharf, S. J., Horn, G. T. & Erlich, H. A. (1986) *Science* **233**, 1076–1078.
9. Sarich, V. M. & Wilson, A. C. (1967) *Science* **158**, 1200–1203.
10. Hasegawa, M., Kishino, H. & Yano, T. J. (1987) *Mol. Evol.* **26**, 132–147.
11. Li, W.-H. & Tanimura, M. (1987) *Nature (London)* **326**, 93–96.
12. Sakoyama, Y., Hong, K.-J., Byun, S. M., Hisajima, H., Ueda, S., Yaoita, Y., Hayashida, H., Miyata, T. & Honjo, T. (1987) *Proc. Natl. Acad. Sci. USA* **84**, 1080–1084.

13. Sibley, C. G. & Ahlquist, J. E. J. (1987) *Mol. Evol.* **26,** 99–121.
14. Pilbeam, D. (1984) *Sci. Am.* **250** (3), 84–96.
15. Li, W.-H., Tanimura, M. & Sharp, P. M. (1987) *Mol. Evol.* **25,** 330–342.
16. Mullis, K. B. & Faloona, F. A. (1987) *Methods Enzymol.* **155,** 335–350.
17. Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A. & Arnheim, N. A. (1985) *Science* **230,** 1350–1354.
18. 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.
19. Fisher, S. G. & Lerman, L. S. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 1579–1583.
20. WHO Nomenclature Committee (1988) *Immunogenetics* **28,** 391–398.
21. Farris, J. S. (1970) *Syst. Zool.* **19,** 83–92.
22. Fitch, W. M. (1971) *Syst. Zool.* **20,** 406–415.
23. Brown, J. H., Jardetzky, T., Saper, M. A., Samraoui, B., Bjorkman, P. J. & Wiley, D. C. (1988) *Nature (London)* **332,** 845–850.
24. Stewart, C.-B., Schilling, J. W. & Wilson, A. C. (1987) *Nature (London)* **330,** 401–404.
25. Nei, M. (1987) *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York).
26. Nathenson, S. G., Geliebter, J., Pfaffenbach, G. M. & Zeff, R. A. (1986) *Annu. Rev. Immunol.* **4,** 471–502.
27. Arden, B. & Klein, J. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 2342–2346.
28. Howard, J. C. (1988) *Nature (London)* **332,** 588–590.
29. McConnell, T. J., Talbot, W. S., McIndoe, R. A. & Wakeland, E. K. (1988) *Nature (London)* **332,** 651–654.
30. Klein, J. (1987) *Hum. Immunol.* **19,** 155–162.
31. Figueroa, F., Günther, E. & Klein, J. (1988) *Nature (London)* **335,** 265–268.
32. Lawlor, D. A., Ward, F. E., Ennis, P. D., Jackson, A. P. & Parham, P. (1988) *Nature (London)* **335,** 268–271.
33. Braunstein, N. S. & Germain, R. N. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 2921–2925.
34. Germain, R. N., Bentley, D. M. & Quill, H. (1985) *Cell* **43,** 233–242.
35. Fan, W., Kasahara, M., Gutknecht, J., Klein, D., Mayer, W. E., Jonker, M. & Klein, J. (1989) *J. Hum. Immunol.* **26,** 107–121.
36. Hughes, A. & Nei, M. (1988) *Nature (London)* **335,** 167–170.
37. Kwok, W. W., Thurtle, P. & Nepom, G. T. (1989) *J. Immunol.* **143,** in press.