Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 15285–15288, December 1996
Genetics

Mutation rate varies among alleles at a microsatellite locus: Phylogenetic evidence

LI JIN[†], CLAUDIA MACAUBAS[‡], JOACHIM HALLMAYER[†‡], AKINORI KIMURA[§], AND EMMANUEL MIGNOT[‡¶]

[†]Department of Genetics, and [‡]Center for Narcolepsy, Department of Psychiatry, Stanford University, Stanford, CA 94305; and [§]Department of Tissue Physiology, Tokyo Medical and Dental University, Tokyo, 101, Japan

Communicated by C. C. Tan, Fudan University, Shanghai, People's Republic of China, September 4, 1996 (received for review June 5, 1996)

ABSTRACT The understanding of the mutational mechanism that generates high levels of variation at microsatellite loci lags far behind the application of these genetic markers. A phylogenetic approach was developed to study the pattern and rate of mutations at a dinucleotide microsatellite locus tightly linked to HLA-DQB1 (DQCAR). A random Japanese population (n = 129) and a collection of multiethnic samples (n = 941) were typed at the DQB1 and DQCAR loci. The phylogeny of DQB1 alleles was then reconstructed and DQCAR alleles were superimposed onto the phylogeny. This approach allowed us to group DQCAR alleles that share a common ancestor. The results indicated that the DQCAR mutation rate varies drastically among alleles within this single microsatellite locus. Some DQCAR alleles never mutated during a long period of evolutionary time. Sequencing of representative DQCAR alleles showed that these alleles lost their ability to mutate because of nucleotide substitutions that shorten the length of uninterrupted CA repeat arrays; in contrast, all mutating alleles had relatively longer perfect CA repeat sequences.

Microsatellites, which are abundant in the human genome, are highly polymorphic due to allelic variation in the number of repeat units of 2–5 base pairs. These genetic markers are widely used in human genetics, although an understanding of the mutational mechanism that generates such a high level of variation lags far behind their applications. The high similarity in allele sizes at each locus inspired the hypothesis that stepwise mutation mechanisms through replication slippage might be involved (1–4). The direct knowledge of spontaneous mutation and the estimation of mutation rates in microsatellites are largely due to the contribution of large scale pedigree studies in the search for disease genes (5–9). Such studies have shown that more than 90% of mutations result in the expansion or contraction of the alleles by a single repeat unit (2 to 4 bp). These studies also estimated that the mutation rates of microsatellites studied range from 1.2×10^{-4} to 1.5×10^{-2}. Comparison of allele sequences revealed very complex mutation patterns (10–12). However, the laborious pedigree studies prevent one from observing enough mutations for each locus and consequently all the conclusions drawn are based on a large collection of loci.

Population genetic studies of microsatellite loci have been fruitful in revealing possible pattern of mutations. It has been shown that a simple stepwise mutation model (SMM; ref. 13), which is intuitively compatible with replication slippage mechanism, can describe the behaviors of many microsatellite loci in the populations (14–16). An extended SMM that allows a few big changes in repeat number (multistep SMM) may be more suitable for microsatellites (16, 17). However, since only simple summary statistics were used in the above studies, characterization of individual locus was not possible. Furthermore, the pattern and rate of mutations at microsatellites may vary among loci (14, 18). The understanding of the pattern and rate of mutations is very relevant to the applications of those genetic markers in evolutionary studies as well as in gene mapping studies (19–22).

Phylogenetic approaches reconstruct evolutionary relationships not only among taxa at various levels but also among genes as well as alleles from a single locus. A phylogeny usually spans over a much larger number of generations or meiotic events than a pedigree. Consequently, phylogenetic approaches allow the study of low frequency genetic events such as mutations or recombinations using a relatively smaller sample size. The objective of this work is to demonstrate the utility of this approach in studying the tempo and mode of mutations at a microsatellite locus while the principles used have much broader applications for problems of similar nature.

A CA-repeat microsatellite locus tightly linked to HLA DQB1 locus (DQCAR; ref. 23) was recently characterized in several populations (24). Since DQB1 can be typed relatively easily, it provides a unique opportunity to reconstruct the evolutionary history of DQCAR and consequently to study the pattern and rate of DQCAR mutations by identifying common ancestors of DQCAR alleles. By typing random Japanese samples and a collection of multiethnic samples at both DQB1 and DQCAR, we could demonstrate that mutation rates vary drastically among alleles of a single microsatellite locus. The mechanism of such variation is discussed in this report.

MATERIALS AND METHODS

A Japanese population of 129 unrelated individuals was collected and typed at DRB1, DQA1, and DQB1. A collection of 941 samples from various ethnic backgrounds (Japanese, Papua New Guinean, African-American, and Caucasians including cell line and patients with various autoimmune diseases) were also typed at DRB1, DQA1, and DQB1 in several clinical laboratories and in our lab. All the samples selected for this study were oligotyped by using PCR–sequence specific oligonucleotide probe or related method at DQB1 (see refs. 24 and 25 for detailed descriptions). Typing at DQCAR locus was performed in our lab (24).

DQB1-DQCAR haplotypes were inferred using three different methods. First, haplotypes were inferred based on known associations established from homozygotes and previously published papers (24). An expectation maximization algorithm developed by Excoffier and Slatkin (26) was used to estimate haplotype frequencies. A new haplotyping algorithm based on parsimony principle by minimizing the number of conflicting inferences with exhaustive search was also used (L.J., unpublished work).

The phylogeny of DQB1 alleles was reconstructed based upon the aligned complete coding sequences obtained from the European Molecular Biology Laboratory Data Library.

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.

[¶]To whom reprint requests should be addressed at: Center for Narcolepsy, 701 Welch Road, Suite 2226, Stanford, CA 94304.

Several measures of genetic distance including Kimura's two-parameter model (27) were used in the reconstruction of distance matrix phylogenies (see refs. 28 and 29 for the definition of genetic distance measures and the detailed description of phylogeny reconstruction). The maximum parsimony method was also used (30). The tree was rooted by including pig and horse DQB sequences. MEGA (29) was used in this analysis.

RESULTS

Two Major Groups of DQB1 Alleles Were Found. A phylogeny of 14 DQB1 alleles presented in our samples was reconstructed based upon the aligned complete coding sequences. Fig. 1 presents a Neighbor-joining tree (31) using Kimura's two-parameter model (27). Several other measures of genetic distance were also used, and all gave identical results in terms of the topology of phylogeny using the Neighbor-joining method. The maximum parsimony tree displayed a slightly different topology which became identical with the Neighbor-joining tree shown in Fig. 1 once peptide-binding sites (PBSs) were removed from the analysis. The discrepancy of the phylogeny with and without PBS is probably due to either the variation of substitution rate among nucleotide sites driven by selection variation (32) or the very frequent gene conversion (33). Interestingly, DQB1 alleles were grouped into two major clusters: DQB1*02 (0201, 0202), DQB1*03 (0301, 0302, 03032), and DQB1*04 (0401, 0402) formed the non-DQ1 group; and DQB1*05 (0501, 0502, 05031) and DQB1*06 (06011, 0602, 0603, 0604) formed another cluster (DQ1). The high bootstrapping values shown on internal branches indicate the reliability of the phylogeny (see ref. 29 for detailed explanation of bootstrapping procedure).

The Level of DQCAR Variation Differs Greatly Between the Two Groups of DQB1 Alleles. DQB1-DQCAR haplotypes were inferred using three different methods. The haplotypes were first inferred manually based on (*i*) the haplotypes of homozygote cell lines, (*ii*) known associations established from homozygotes at one of the two loci (DQB1 or DQCAR), and (*iii*) pedigree information for some individuals. Two independent inferrences generated identical results. A maximum likelihood algorithm developed by Excoffier and Slatkin (23) was used to estimate haplotype frequencies. Only those haplotypes with non-zero frequencies were accepted. A new haplotyping algorithm based on parsimony principle minimizing the number of conflicting inferences with exhaustive search was also used (L.J., unpublished work). Identical haplotypes were obtained using all three different methods.

Table 1. Haplotype frequencies of DQB1-DQCAR in Japanese samples

DQB1 alleles	DQCAR allele size				Total
DQB1*0201	99(1)				1
DQB1*0202	113(1)				1
DQB1*0301	113(3)	117(14)	121(12)	123(4)	33
DQB1*0302	109(1)	111(11)	113(10)		22
DQB1*03032	111(1)	115(30)	117(2)		33
DQB1*0401	113(16)	115(9)	117(4)		29
DQB1*0402	113(3)	115(2)	117(4)		9
DQB1*0501	103(22)				22
DQB1*0502	103(3)				3
DQB1*05031	107(5)				5
DQB1*0602	103(17)				17
DQB1*0603	103(2)				2
DQB1*0604	103(21)				21
DQB1*06011	107(60)				60

Numbers in parentheses are numbers of chromosomes.

The Japanese DQB1-DQCAR haplotypes (number of chromosomes in parenthesis) are presented in Table 1. The haplotype data for the mixed population including both 941 multiethnic samples and 129 Japanese samples are shown in Fig. 2. The alleles observed only once in all 2140 chromosomes are indicated by parentheses (Table 2). The number of each DQB1 allele is listed in the last column of Fig. 2. For both the Japanese and the mixed population samples, the combined observed frequencies for DQ1 and non-DQ1 alleles were similar. However, the level of DQCAR variation differs greatly between the two groups of DQB1 alleles. In both populations, the numbers of DQCAR alleles in non-DQ1 lineages were larger than those found in the DQ1 lineages, thus suggesting that DQCAR alleles associated with non-DQ1 alleles might have much higher mutation rates than those associated with DQ1 alleles. Furthermore, the DQCAR alleles found in non-DQ1 lineages tended to have larger fragment sizes (109–125 bp) than those observed in DQ1 lineages with the exception of DQB1*0201. In contrast, most of the DQCAR alleles in the DQ1 group were monomorphic with a 103-bp size (24).

The number of DQCAR alleles observed in our population for a given DQB1 lineage should be the result of several factors: (*i*) The mutation rate of each individual DQCAR allele, (*ii*) the age of the DQB1 lineage studied, and (*iii*) the

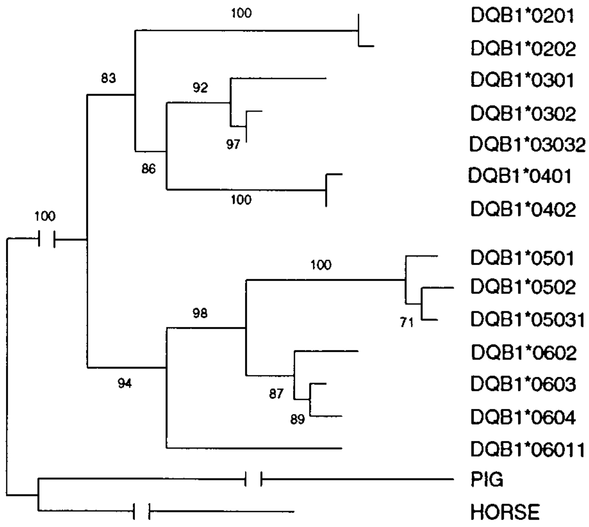


FIG. 1. Neighbor-joining tree of DQB1 based on Kimura's two parameter distance. The numbers are bootstraping values (in percentage) with 500 replications.

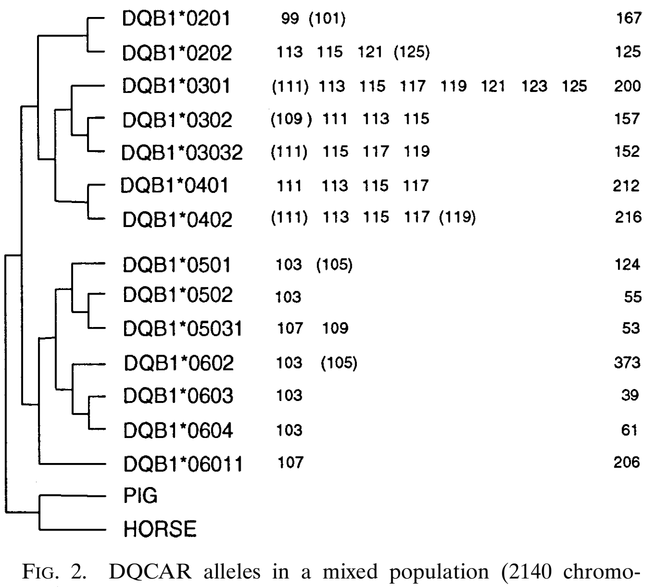

DQB1*0201 99 (101) 167
DQB1*0202 113 115 121 (125) 125
DQB1*0301 (111) 113 115 117 119 121 123 125 200
DQB1*0302 (109) 111 113 115 157
DQB1*03032 (111) 115 117 119 152
DQB1*0401 111 113 115 117 212
DQB1*0402 (111) 113 115 117 (119) 216
DQB1*0501 103 (105) 124
DQB1*0502 103 55
DQB1*05031 107 109 53
DQB1*0602 103 (105) 373
DQB1*0603 103 39
DQB1*0604 103 61
DQB1*06011 107 206
PIG
HORSE

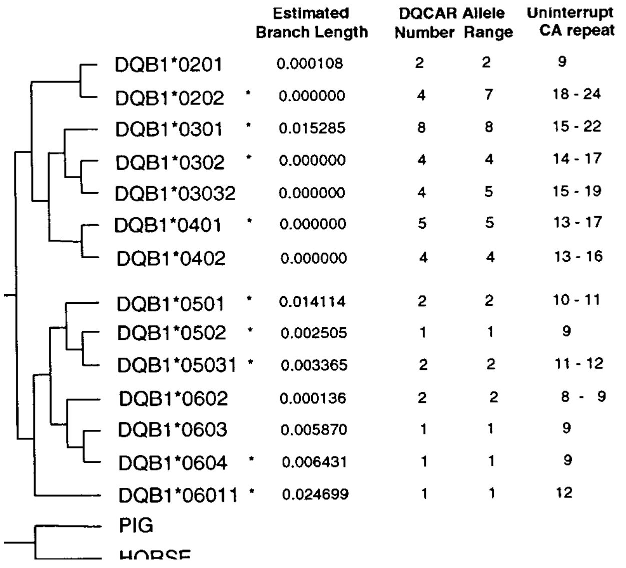
FIG. 2. DQCAR alleles in a mixed population (2140 chromosomes).

effective number of individuals in the population bearing the DQB1 allele. The relative importance of these factors can be estimated by studying the properties of the phylogeny and the frequency distribution of DQB1 allele in various populations.

Do DQCAR Alleles Associated with Specific DQB1 Lineage Share a Common Ancestor? The DQCAR alleles associated with a given DQB1 allele share a unique common ancestral microsatellite allele right after the DQB1 lineage emerged in the population. This becomes only evident if it can be shown that the "new" DQB1 allele contains specific mutation(s) unique to this lineage. A maximum parsimony principle (30) was used to infer the mutations for each external lineages of the DQB1 phylogeny based upon the sequences of oligonucleotide (24) used in genotyping. A DQB1 lineage was considered as a new derivative if it had a unique mutation that was not present in other lineages. Sites that have to be explained by multiple and/or recurrent mutations (most probably due to gene conversion) were excluded from the analysis. An asterisk is added to each of the newly derived DQB1 lineage in Fig. 3. Note that the phylogeny was reconstructed (see Fig. 1) based on the sequences of the complete coding region while the samples were typed by PCR-SSOP oligotyping which includes only a few sites within the DQB1 gene (mostly in the second exon).

The age of DQB1 lineage-specific mutations can be estimated by the length of external lineages. The branch lengths of these lineages were estimated using synonymous substitution rate (34) based on the full-length sequences but not those of oligonucleotide. These data are presented in Fig. 3 along with the number and the range of DQCAR allele sizes for each DQB1 lineage observed in the multi-ethnic population. The number of allele is the total number of alleles associated with each DQB1 lineage. The range of allele is the number of possible CA repeat units between the minimum and maximum observed alleles (including both). For example, if the sizes of allele associated with DQB1*0202 range from 113 bp to 125 bp, the range of allele observed in this DQB1 lineage is (125–113)/2 + 1 = 7. Both the number and the range of alleles associated with a DQB1 lineage reflect the level of variation at the DQCAR locus (17).

A careful observation of the phylogeny displayed in Fig. 3 clearly demonstrate that the number of DQCAR allele observed in each individual DQB1 branch does not correlate with the estimated branch length. In all cases, the number and the range of DQCAR alleles observed in the non-DQ1 lineages are much larger than those associated with DQ1 lineages independent of branch length and of the number of chromosomes tested with each DQB1 subtype. This was evident not only for alleles that were found frequently within a specific population (e.g., DQB1*06011 in Japanese, DQB1*02 in Caucasians) but also for alleles with high frequency across various ethnic groups (DQB1*03, DQB1*04, DQB1*0602, and DQB1*0604; see ref. 35 for DQB1 allele frequencies across worldwide populations). For example, DQB1*06011 is featured with an extremely long lineage and a very large number of individuals compared with others but it is still monomorphic at the DQCAR locus. Similarly, DQB1*0602 is generally monomorphic in a very large number of individuals across all populations. In contrast, several DQB1 lineages such as DQB1*0202, DQB1*0302, and DQB1*0401 have almost negligible branch lengths, yet they have much larger numbers of DQCAR alleles. Therefore, the variation of effective numbers of chromosomes bearing DQB1 lineages does not necessarily contribute to the much larger variation of DQCAR alleles in non-DQ lineages.



		Estimated Branch Length	DQCAR Allele Number	DQCAR Allele Range	Uninterrupt CA repeat
DQB1*0201		0.000108	2	2	9
DQB1*0202	*	0.000000	4	7	18 - 24
DQB1*0301	*	0.015285	8	8	15 - 22
DQB1*0302	*	0.000000	4	4	14 - 17
DQB1*03032		0.000000	4	5	15 - 19
DQB1*0401	*	0.000000	5	5	13 - 17
DQB1*0402		0.000000	4	4	13 - 16
DQB1*0501	*	0.014114	2	2	10 - 11
DQB1*0502	*	0.002505	1	1	9
DQB1*05031	*	0.003365	2	2	11 - 12
DQB1*0602		0.000136	2	2	8 - 9
DQB1*0603		0.005870	1	1	9
DQB1*0604	*	0.006431	1	1	9
DQB1*06011	*	0.024699	1	1	12
PIG					
HORSE					

FIG. 3. Estimated branch length for each DQB1 lineage. Number, size range, and the number of uninterrupted CA repeats of DQCAR alleles carried by each DQB1 lineage.

Mutation Rates Vary Among DQCAR Alleles. The above observation suggests that DQCAR alleles with larger fragment sizes in non-DQ1 lineages (109–125 bp) have higher mutation rates than DQ1 DQCAR alleles with small fragment sizes (103 bp). The only exception is DQB1*0201, a non-DQ1 allele, which shows low DQCAR variation in a large number of chromosomes in spite of its position in the tree. In this lineage, however, the allele size of DQCAR (99 bp) is the smallest amongst all samples. This is thus in fact consistent with the hypothesis that microsatellite alleles with a larger number of repeats tend to have higher mutation rates (4).

At least one DQCAR allele for each DQB1 lineage was subcloned and sequenced. All non-DQ1 DQCAR alleles were found to share almost identical flanking sequences while those for DQ1 were also identical but different from non-DQ1 DQCAR sequences (C.M., unpublished work). Several nucleotide substitutions were observed in DQCAR allele sequences associated with DQ1 lineages, and one of them occurred in the middle of the CA repeat array disrupting the CA repeat structure. The number of uninterrupted CA repeats for each DQB1 lineage is presented in Fig. 3. Non-DQ1 DQCAR alleles do have a larger number of CA repeats than those bearing DQ1 haplotypes. Interestingly, the number of CA repeats at DQB1*0201 is 9, similar to DQ1 DQCAR alleles, although its length (99 bp) is shorter than the latter (103 bp).

DISCUSSION

In this report, we demonstrate that DQB1 alleles can be classified into two major groups: DQ1 and non-DQ1. The number of DQCAR alleles associated with these two DQB1 groups was also found to vary drastically, with non-DQ1 lineages being very variable and DQ1 lineages being almost monomorphic. Further analysis indicated that this difference correlates mostly with the number of uninterrupted CA repeat sequences within the microsatellite rather than other population genetic factors such as the age of DQB1 lineages, effective population sizes, or recent population expansion.

The DQCAR allele sizes observed within individual DQB1 lineages were very close to each other and often differed by increments of 2 bp, indicating that replication slippage is a reasonable hypothetical mechanism underlying size polymorphism at the level of the DQCAR locus. This was especially obvious for the low mutating lineages such as DQB1*0201, DQB1*0501, DQB1*05031, and DQB1*0602 where only two neighboring DQCAR alleles (differing by 2 bp) were observed suggesting single step replication slippage events.

Major histocompatibility complex alleles are well known to be subject to balance selection (30), thus tending to have much longer lineages than neutral markers. It is expected therefore to be an ideal system to study low frequency genetic events. In spite of this, three out of four informative lineages in non-DQ1 show negligible numbers of synonymous substitutions, indicating one might have to study longer sequences. The presence of natural selection could be a problem since it would make it difficult to compare traits across DQB1 lineages if the intensity of the selection among lineages are different. However, there is no reason to believe that DQCAR alleles associated with a single DQB1 lineage are subject to different level of natural selection. Therefore, the conclusions in this report that were mostly based upon the number of alleles among lineages would not be affected by such variation of selection.

Recombination between DQCAR and the DQB1 gene could be a problem. However, the recombination rate within the HLA class II region is extremely low as indicated by strong linkage disequilibrium (36). Furthermore, we have never observed the most frequent DQ1 specific DQCAR allele (103 bp), which has an extremely high frequency in natural populations in any non-DQ1 sample. We therefore conclude that recombination between DQCAR and DQB1 is negligible.

Many approaches in human genetics rely solely on pedigree data. A large number of individuals and pedigrees have to be typed to study low frequency events such as mutations and recombinations since pedigrees only span over a few generations. In contrast, a phylogenetic approach includes events of interest that have accumulated over hundreds or thousands of generations. In this report, we demonstrated the utility of the approach in studying the mutational mechanisms of a microsatellite locus. The principle employed, we believe, could have much broader applications for problems of similar nature.

This work was supported by National Institute of Neurological Disorders and Stroke Grant NS33797 to E.M. and by National Institutes of Health Training Grant T32-GS08404 to L.J. We thank Drs. P. Underhill and P. Oefner for the suggestions regarding cloning and sequencing strategies, Dr. L. Excoffier for providing program EXHAPFRE, and Drs. N. Takahata and T. Ota for technical suggestions on phylogeny reconstruction of DQB1 alleles.

1. Tautz, D. & Renz, M. (1984) *Nucleic Acids Res.* **12,** 4127–4138.
2. Levinson, G. & Gutman, G. (1987) *Mol. Biol. Evol.* **4,** 203–221.
3. Stephen, W. (1989) *Mol. Biol. Evol.* **6,** 198–212.
4. Weber, J. (1990) *Genomics* **7,** 524–530.
5. Kwiatkowski, D., Henske, E., Weimer, K., Ozelius, L., Gusella, J. & Haines, J. (1992) *Genomics* **12,** 229–240.
6. Petrukhin, K., Speer, M., Cayanis, E., Bonaldo, M., Tantravali, U., Soares, M., Fischer, S., Warburton, D., Gilliam, T. & Ott, J. (1993) *Genomics* **15,** 76–85.
7. Bowcock, A., Osborne-Lawrence, S., Barnes, R., Chakravarti, A., Washington, S. & Dunn, C. (1993) *Genomics* **15,** 376–386.
8. Mahtani, M. & Willard, H. (1993) *Hum. Mol. Genet.* **2,** 431–437.
9. Weber, J. & Wong, C. (1993) *Hum. Mol. Genet.* **8,** 1123–1128.
10. Puers, C., Hammond, H. A., Jin, L., Caskey, T. & Schumm, J. W. (1993) *Am. J. Hum. Genet.* **53,** 953–958.
11. Garza, J. C. & Freimer, N. B. (1996) *Genome Res.* **6,** 211–217.
12. Garza, J. C., Slatkin, M. & Freimer, N. B. (1995) *Mol. Biol. Evol.* **4,** 594–603.
13. Ohta, T. & Kimura, M. (1973) *Genet. Res.* **22,** 201–204.
14. Edwards, A., Hammond, H., Jin, L., Caskey, C. & Chakraborty, R. (1992) *Genomics* **12,** 241–253.
15. Valdes, A., Slatkin, M. & Freimer, N. (1993) *Genetics* **133,** 737–749.
16. Shriver, M., Jin, L., Chakraborty, R. & Boerwinkle, E. (1993) *Genetics* **134,** 983–993.
17. Di Rienzo, A., Peterson, A., Garza, J., Valdes, A. & Slatkin, M. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 3166–3170.
18. Jin, L., Zhong, Y., Shriver, M., Deka, R. & Chakraborty, R. (1994) *Am. J. Hum. Genet.* **55,** Suppl., A39.
19. Shriver, M., Jin, L., Chakraborty, R. & Boerwinkle, R. (1993) *Am. J. Hum. Genet.* **53,** Suppl., 858.
20. Slatkin, M. (1995) *Genetics* **139,** 457–462.
21. Goldstein, D., Linares, A., Cavalli-Sforza, L. & Feldman, M. (1995) *Genetics* **139,** 463–471.
22. Shriver, M., Jin, L., Boerwinkle, E., Deka, R., Ferrell, E. & Chakraborty, R. (1995) *Mol. Biol. Evol.* **12,** 914–920.
23. Satyanarayana, K. & Strominger, J. (1992) *Immnogenetics* **35,** 235–240.
24. Macaubas, C., Hallmayer, J., Kalil, J., Kimura, A., Yasunaga, S., Grumet, F. & Mignot, E. (1995) *Hum. Immnol.* **42,** 209–220.
25. Kimura A., Dong, R., Harada, H. & Sasazuki, T. (1992) *Tissue Antigens* **40,** 5–12.
26. Excoffier, L. & Slatkin, M. (1995) *Mol. Biol. Evol.* **12,** 921–927.
27. Kimura, M. (1980) *J. Mol. Evol.* **16,** 111–120.
28. Nei, M. (1987) *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York).
29. Kumar, S., Tamura, R. & Nei, M. (1993) *MEGA, Molecular Evolutionary Genetics Analysis* (Pennsylvania State Univ., University Park, PA), Version 1.0.
30. Fitch, W. (1981) *J. Mol. Evol.* **18,** 30–37.
31. Saitou, N. & Nei, M. (1987) *Mol. Biol. Evol.* **4,** 406–425.
32. Hughes, A. & Nei, M. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 958–962.
33. Zangenberg, G., Huang, M.-M., Arnheim, N. & Erlich, H. (1995) *Nat. Genet.* **10,** 407–414.
34. Nei, M. & Gojobori, T. (1986) *Mol. Biol. Evol.* **3,** 418–426.
35. Tsuji, K., Aizawa, M. & Sasazuki, T. (1992) *Proceedings of the Eleventh International Histocompatibility Workshop and Conference* (Oxford Univ. Press, Oxford).
36. Martin, M., Mann, D. & Carrington, M. (1995) *Hum. Mol. Genet.* **4,** 423–428.