The expected distribution of RAPD bands

Weissing, F.J.; Velterop, O.

Published in:
Molecular Biology and Evolution

Publication date:
1993

Prospects for Estimating Nucleotide Divergence with RAPDs[1]

Andrew G. Clark and Caroline M. S. Lanigan†*

*Institute of Molecular Evolutionary Genetics, Department of Biology, Pennsylvania State University; and †Comparative Genetics Subunit, California Regional Primate Research Center, University of California

The technique of random amplification of polymorphic DNA (RAPD), which is simply polymerase chain reaction (PCR) amplification of genomic DNA by a single short oligonucleotide primer, produces complex patterns of anonymous polymorphic DNA fragments. The information provided by these banding patterns has proved to be of great utility for mapping and for verification of identity of bacterial strains. Here we consider whether the degree of similarity of the banding patterns can be used to estimate nucleotide diversity and nucleotide divergence. With haploid data, fragments generated by RAPD-PCR can be treated in a fashion very similar to that for restriction-fragment data. Amplification of diploid samples, on the other hand, requires consideration of the fact that presence of a band is dominant to absence of the band. After describing a method for estimating nucleotide divergence on the basis of diploid samples, we summarize the restrictions and criteria that must be met when RAPD data are used for estimating population genetic parameters.

Introduction

Random amplification of polymorphic DNA (RAPD) by the polymerase chain reaction (PCR), or RAPD-PCR, is a means of rapidly detecting polymorphisms for genetic mapping and strain identification (Welsh and McClelland 1990; Williams et al. 1990). The method applies the PCR with a single short oligonucleotide primer, randomly amplifying short fragments of genomic DNA, which are size-fractionated by agarose gel electrophoresis. The method has considerable appeal because it is generally faster and less expensive than any previous method for detecting DNA sequence variation. The fact that RAPDs survey numerous loci in the genome makes the method particularly attractive for analysis of genetic distance and phylogeny reconstruction. There is need, however, to develop quantitative measures of genetic similarity based on observed RAPD patterns.

The RAPD method is useful in genetic analysis only if variation in banding patterns represents allelic segregation at independent loci. Polymorphism is detected as band presence versus absence and may be caused either by failure to prime a site in some individuals because of nucleotide sequence differences or by insertions or deletions in the fragment between two conserved primer sites. True allelic segregation may be confused with intermittent PCR artifacts (Riedy et al. 1992), unless additional

1. Key words: RAPDs, PCR, nucleotide diversity, nucleotide divergence.

Address for correspondence and reprints: Andrew G. Clark, Institute of Molecular Evolutionary Genetics, Department of Biology, Pennsylvania State University, University Park, Pennsylvania 16802.

Mol. Biol. Evol. 10(5):1096–1111. 1993.

0737-4038/93/1005-0012$02.00

genetic analysis is performed. Other technical problems and a number of experimental solutions have been presented by Hadrys et al. (1992) and Lanigan (1992). Pedigree-structured data make it possible to identify bands that exhibit Mendelian segregation. In addition, RAPDs provide a powerful means of assembling a map of anonymous segregating markers from an array of F_2 individuals derived from a pair of inbred lines, because only markers that exhibit good genetic segregation will provide sensible genetic map information (Martin et al. 1991).

Estimation of the degree of nucleotide sequence divergence between two or more individuals is a problem that frequently arises in evolutionary biology. The purpose of this note is to point out some criteria that must be met before RAPDs can be used for this purpose and to illustrate how RAPDs can be used when these criteria are satisfied.

Expected Number of RAPD Bands

Let us begin with the assumption that the primers hybridize to all sites in the genome that precisely match in sequence and that one or more base mismatches at the primer site precludes hybridization. Hypothetically, assume that a fragment is amplified in all cases in which two successive primer sites are located on complementary strands in opposite orientation. This is equivalent to assuming that, when primer sites are nested, only the smallest of the possible fragments is amplified, and in practice this appears to be true (Williams et al., in press). Under the assumption that the nucleotide sequence is random and unstructured, the probability of matching a primer at a site is independent of the proximity of other matches, provided that the primer sequence is non-self-overlapping. This means that the occurrence of matches can be modeled as a discrete renewal process, with an exponential distribution of distances between matches and with a Poisson distribution of the number of matches (Feller 1968, pp. 1–11). If a is the probability that any given n-nucleotide run in the genomic DNA matches the n-nucleotide primer, then, if successive bases are statistically independent, a is the product of the genomic frequencies of the bases in the primer. If the bases were equally frequent, $a = (1/4)^n$. The expected distribution of the interval between such primer sites on one strand of the DNA is exponential with parameter a. The same distribution applies to the opposite strand. Amplification occurs only when successive primer sites are of opposite orientation, and renewal theory shows that the expected fragment length is $1/(2a)$ (see the Appendix). The distribution of number of primer sites is Poisson with mean Ca, where C is the number of nucleotides in the genome. Because adjacent primers must be of opposite orientation, and are therefore not independent, the number of fragments amplified does not have a Poisson distribution. Derivation of the distribution of the number of fragments is presented in the Appendix, and its expectation is $Ca/2$. If only fragments between sizes s_1 and s_2 are recovered, the number of visible bands is approximately $CaV/2$, where,

$$V = \sum_{x=s_1}^{s_2} 2a(1-2a)^{x-1} = (1-2a)^{s_1-1} - (1-2a)^{s_2}. \qquad (1)$$

If the primer sequence can overlap with itself, a needs to be corrected in a manner analogous to that described by Waterman (1983).

It is instructive to compare the theoretical prediction to empirical observations

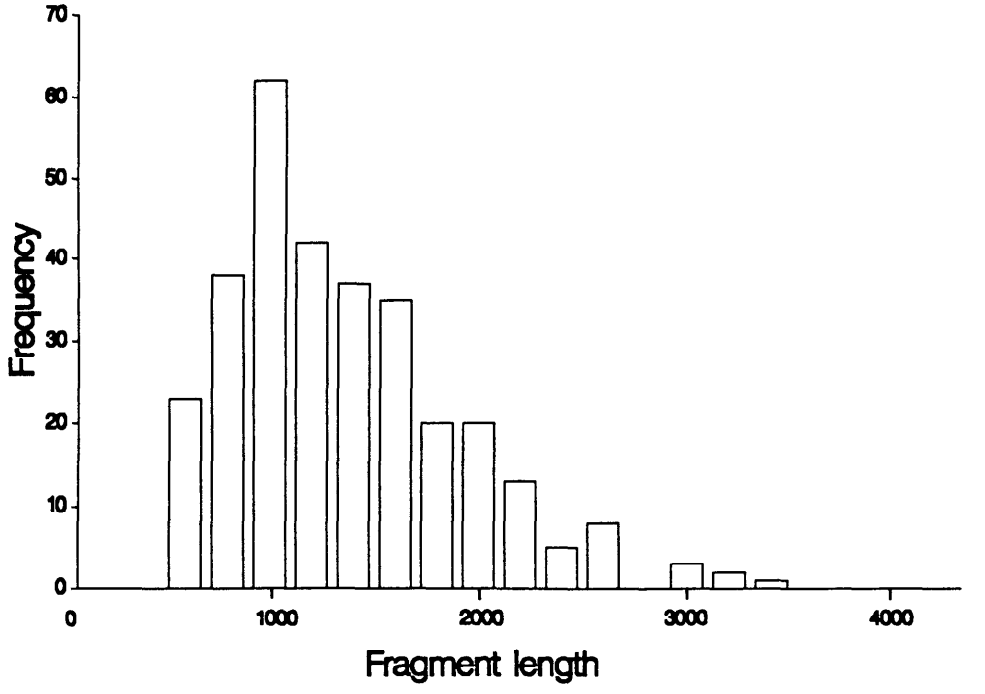
of RAPD banding patterns. In the case of RAPD analysis of macaques, whose haploid genome is $\sim 3 \times 10^9$ nucleotides (nt), the expected fragment length generated by a 10-nt primer is $\frac{1}{2} \times 4^{10} = 524{,}288$ nt, and the expected number of RAPD bands that would be generated if PCR amplified all possible bands is $3 \times 10^9 / 2{,}097{,}152 = 1{,}430.5$. The fraction of all bands expected to be between 500–3,500 nt in length is $V = 0.0057$, so the expected number of fragments in this size interval is 8.16. This is the expected number of bands for a haploid genome, and a diploid genome will produce more bands, because of polymorphism. The fact that eukaryotic genomes are partitioned into chromosomes has little effect on the expected number of fragments, because detectable fragments are so small in comparison with chromosome lengths that the chance that a detectable fragment is interrupted by a chromosome end is very small. In a survey of DNA polymorphism in a sample of eight macaques, Lanigan (1992) applied RAPD-PCR with 22 primers each of length 10 nt and observed a mean of 11 bands/individual (fig. 1).

There are a number of reasons why the observed number of bands is greater than the theoretical prediction, and it is unlikely that further pursuit of this line of theory will provide the best estimator of nucleotide divergence. DNA sequences exhibit nonrandomness at nearly all levels, from dinucleotides (Bulmer 1986) to scales of 10s of kilobases (Bernardi et al. 1985; Bernardi 1989). Some bands may also be generated when mispriming results in amplification at sites that do not perfectly match the primer sequence. The observation that many 10-nt primers produce five or more RAPD bands from bacterial genomic DNA suggests that either mispriming occurs frequently or bacterial genomes are highly structured (T. Whittam, personal communication). We conclude that an estimate of divergence must rely not on absolute counts of bands but, rather, on the proportion of bands that are shared by two (or more) samples.

Estimation of Nucleotide Divergence

RAPD data can be treated as analogous to restriction-fragment presence/absence, for estimating nucleotide divergence (number of substitutions per site), if a few assumptions are made. First, we assume that the amplification of a fragment depends strictly on the exact match between the oligonucleotide primer and a site on the genomic DNA. Thus, if one DNA sample amplifies a particular band and another DNA sample does not, we assume that a single nucleotide substitution in a primer site is sufficient to account for the difference. Because multiple hits in a primer site are treated as a single substitution, the analysis is restricted to sequences that have diverged less than $\sim 10\%$. We also make assumptions similar to those of Lynch (1990) for VNTR samples—namely, that fragment sizes are accurately assessed, that the population is sampled at random, that allelism of bands can be determined (e.g., by Southern blotting), that different bands represent independent loci in linkage equilibrium and Hardy-Weinberg genotypic proportions (in the case of diploids), and that the same set of primers is assayed in all individuals. Initially we will ignore insertion/deletion differences. Finally, RAPDs behave as dominant markers (presence of a band is dominant to absence of a band), and the estimates of divergence in diploids must take this into account. For simplicity, we first develop the estimators for the case of haploids (including gametophytes and homozygous lines of diploids), and then we will consider samples from a panmictic diploid population.

Let P be the probability that no mutation has occurred at a primer site since the



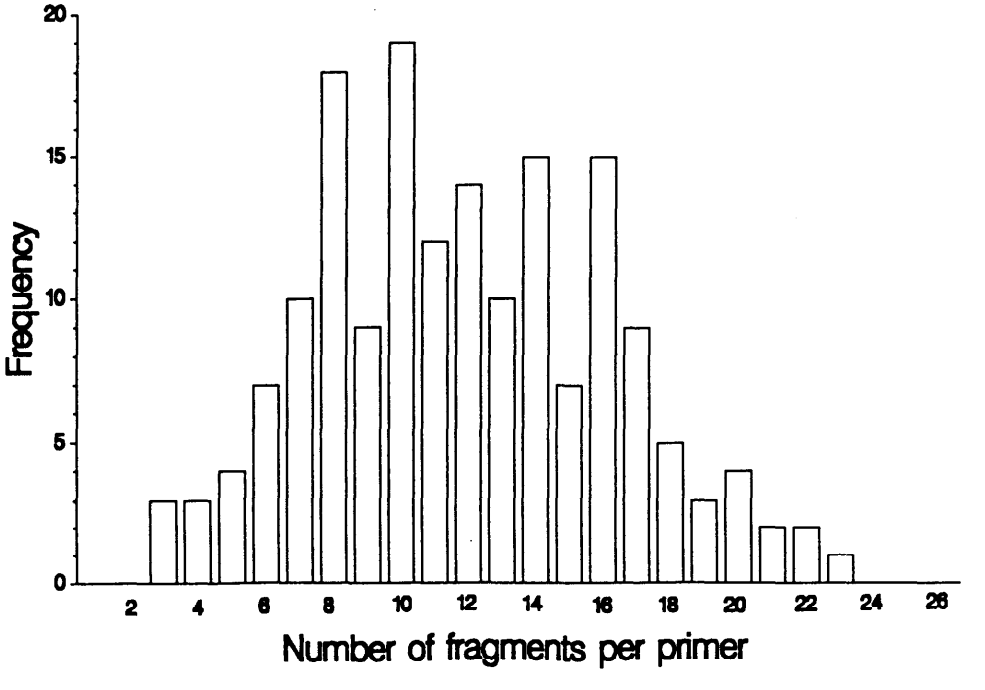


FIG. 1.—*Top,* Observed distribution of fragment lengths obtained from RAPD amplification of genomic DNA from one macaque with 35 different 10-nt primers. *Bottom,* Observed distribution of the number of fragments amplified for each individual × primer combination, for a sample of eight macaques and 22 primers. For experimental details, see the work of Lanigan (1992).

common ancestor of two sequences. If λ is the rate of nucleotide substitution per nucleotide position, t is the time of divergence of the sequences, and r is the length of the primer site, then $P = \exp(-r\lambda t)$. If F is the expected proportion of fragments that remain unchanged, then, according to the reasoning of Nei and Li (1979), in their development of an estimate of nucleotide divergence from restriction-fragment size data,

$$F \approx P^4/(3 - 2P) . \tag{2}$$

From the data, we can tally the counts of bands that are shared by the two individuals (n_{xy}), those present in individual x(n_x), and those present in individual y(n_y) and can estimate F as

$$\hat{F} = 2n_{xy}/(n_x + n_y) . \tag{3}$$

We can estimate $\hat{P}$ from $\hat{F}$ by using the iteration approach suggested by Nei (1987, pp. 106–107):

$$\hat{P} = [\hat{F}(3 - 2\hat{P}_1)]^{1/4} , \tag{4}$$

where $\hat{P}_1$ is the initial trial value of $\hat{P}$ ($F^{1/4}$ is suggested), and the values of $\hat{P}$ obtained from equation (3) are substituted back in as $\hat{P}_1$ until $\hat{P} = \hat{P}_1$. The expected nucleotide divergence is $d = 2\lambda t$, and, since $P = \exp(-r\lambda t)$, this gives an estimate of nucleotide divergence,

$$\hat{d} = -(2/r)\ln(\hat{P}) . \tag{5}$$

The estimate of d can be made as soon as $\hat{F}$, the proportion of bands that are shared, is estimated. Fortunately, the estimate of $\hat{F}$ is independent of the efficiency of PCR in amplifying all the possible target fragments. If only a fraction of the expected fragments are actually detected after RAPD-PCR, then, provided that this fraction is the same for both monomorphic and polymorphic sites, the estimate of the proportion of shared bands will remain valid.

The above estimate of d is the nucleotide divergence for a pair of haploid individuals. If several individuals from each of two populations (or species) are examined, it is possible to estimate the interpopulational nucleotide divergence by following Nei and Miller (1990). If N_X and N_Y genes are samples from population X and Y, respectively, then

$$F_c = \frac{2 \sum_{i,j} n_{X_i Y_j}}{N_Y \sum_i n_{X_i} + N_X \sum_j n_{Y_j}} \tag{6}$$

is calculated, where $n_{X_i Y_j}$ is the number of bands shared between individual i from population X and individual j from population Y, and the sum is over all pairs of individuals drawn one from each population. n_{Xi} is the number of bands seen in individual i of population X, and the sum is weighted by N_Y to make the number of

within- and between-population comparisons the same. Similarly, if many individuals from a single population are examined, it is possible to estimate the nucleotide diversity π by calculating a composite $\hat{F}$ by taking all pairs of individuals.

Because each primer is equally likely to reveal polymorphisms, the variance in nucleotide divergence can be estimated from the variation in estimates obtained from each primer. A convenient numerical means to estimate the variance is to use the jackknife (Efron 1982, pp. 13–19), as was applied to the problem of restriction-site data by Nei and Miller (1990). If $\hat{d}_i$ is the nucleotide divergence between two populations, estimated from all primers except primer i, and if there are m primers used, then

$$V(\hat{d}) = \frac{m-1}{m} \sum_{i=1}^{m} (\hat{d}_i - \bar{d})^2 . \tag{7}$$

If the primers amplify independent sites, then this estimate of variance incorporates the stochastic effects of past historical sampling, as well as the sampling variance from the current sample.

Correcting for Dominance in Diploid Samples

Dominance of RAPDs results in the appearance of greater band sharing among diploid individuals drawn from a panmictic population than is expected among homozygous lines. If the population frequency of the allele lacking a particular band (null allele) is q, then under Hardy-Weinberg conditions the expected frequency of genotypes that exhibit the corresponding band is $1 - q^2$. When $q = 0.1$, 99% of the genotypes will amplify the corresponding band. A sample of 69 genotypes is required before there is a 50% chance of seeing a single case of band absence [the solution to $(1 - q^2)^n = 0.50$]. Dominance will result in an underestimate of nucleotide diversity if the procedure described above is applied without modification.

To correct for dominance, it is necessary to have estimates of band (allele) frequencies, and this requires assuming that the population is in Hardy-Weinberg equilibrium. If z is the frequency of genotypes lacking a particular band, then the large-sample estimate of the null-allele frequency is $q = z^{1/2}$. This is a biased estimator (Lynch and Milligan, submitted), but, although the degree of bias can be substantial when the null allele is rare, in the simulations described below the error in estimates of divergence caused by sampling bias was negligible. The expected heterozygosity under Hardy-Weinberg equilibrium is $2pq$. Define the conditional heterozygosity of band i in an individual from population x, given that band i is observed, as $H_{x(i)} = 2pq/(p^2 + 2pq)$. The values of n_x, n_y, and n_{xy} can be tallied by summing over the $i = 1$–k bands for a pair of individuals from within and between populations x and y:

$$\begin{aligned}
n_x &= \sum_i [4(1 - H_{x(i)})^2 + 4H_{x(i)}(1 - H_{x(i)}) + H_{x(i)}^2] ; \\
n_y &= \sum_i [4(1 - H_{y(i)})^2 + 4H_{y(i)}(1 - H_{y(i)}) + H_{y(i)}^2] ; \\
n_{xy} &= \sum_i [4(1 - H_{x(i)})(1 - H_{y(i)}) + 2(1 - H_{x(i)})H_{y(i)} \\
&\quad + 2H_{x(i)}(1 - H_{y(i)}) + H_{x(i)}H_{y(i)}] .
\end{aligned} \tag{8}$$

The formula for n_{xy} is derived by considering pairs of diploid individuals drawn from each of two populations. For each pair of individuals being compared, we have the four possible interpopulation allelic comparisons. When the pair of individuals being compared are both homozygous for band presence [which occurs in the fraction $(1 - H_{x(i)})(1 - H_{y(i)})$ of the cases in which both individuals exhibit the band], all four allelic comparisons yield identity. In this event, n_{xy} is weighted by 4. When one individual is heterozygous and the other is homozygous for band presence [which occurs with probability $(1 - H_{x(i)})H_{y(i)} + H_{x(i)}(1 - H_{y(i)})$], two of the four allelic comparisons yield an identity, so n_{xy} is incremented by 2. When both individuals are heterozygous (probability $H_{x(i)}H_{y(i)}$), only one allelic comparison yields an identity, so n_{xy} is incremented by 1. Whenever either individual is homozygous for the null allele, n_{xy} is not incremented. n_{xy} is obtained by summing these increments over all scorable bands produced by all primers tested. The calculation of n_x proceeds in a similar fashion, except that now the pairs of diploid individuals are drawn from within population x. Again the four comparisons are made between the pairs of alleles drawn one from each individual. Finally, n_y is calculated by considering pairs of diploid individuals drawn from within population y.

After the weighted values of n_{xy}, n_x, and n_y are tallied for all bands and pairs of individuals, F and d are calculated from equations (3)–(6). The jackknife procedure described above can be applied to estimate $V(d)$. The approach of nucleotide counting (Nei and Tajima 1983) can also be applied after the weighted values of n_{xy}, n_x, and n_y are calculated. Note that parsimony methods cannot be applied directly to RAPD data from diploids, because individuals homozygous for band presence are not distinguished from heterozygotes.

The estimator for d is only valid for $d < 0.10$, in part because no consideration for either multiple substitutions or back mutations was made. Also, because the probability of a mismatch at a primer site is $\sim 1 - (1 - d)^r$, most sites appear to mismatch at a relatively low degree of divergence. Because a number of assumptions and approximations were used to derive these estimators, it is necessary to verify the methods by computer simulation.

Simulations and Model Verification

To test the accuracy of the fragment-sharing approach for estimating nucleotide divergence on the basis of RAPD data, simulations for both the haploid and diploid cases were performed. In the diploid case, it is necessary to consider the frequency of bands in estimating divergence, and for this reason the genealogy of alleles became important. For both the haploid and diploid simulations, a single panmictic population was divided into two equal parts at time 0, and subsequently no migration was allowed. A neutral gene genealogy was generated by using Takahata and Nei's (1985) theory, and mutations were distributed along the branches of the tree by following the Poisson distribution. The parameter $4N\mu$ (where N is the effective population size and μ is the neutral mutation rate per nucleotide site) was set to 0.005 for all simulations, and sample sizes (s) of 10 and 50 individuals/population were generated. Divergence times corresponding to mean sequence divergence of 1%, 2%, . . . , 10% were simulated. This process generated nucleotide sequences of length 1,200 nt, which were scanned for matches with an array of 10 random oligonucleotide primers each of length 4 nt. Resulting fragments of all lengths were considered, and vectors of presence/absence of ~50 bands/population (when $s = 50$) were tallied. For each divergence time, 20

replicate pairs of populations were simulated, and the nucleotide divergence was estimated by following equations (3)–(6). Sample standard deviations were calculated to measure the variation among these replicate samples.

The simulations allow comparison of the true nucleotide divergence to the nucleotide divergence estimated by RAPD band sharing. In the case of haploids, if the nucleotide divergence was less than ~10%, the estimator based on RAPD patterns provides a reasonably good measure of divergence (table 1 and fig. 2). A sample of 50 genomes provides somewhat lower error than does a sample of 10, but samples >50 probably do not improve the estimate significantly.

The diploid case was simulated in a similar fashion. For a sample of s individuals, $2s$ haplotypes were generated in each of two diverging populations. These were then combined in pairs to form s diploid genotypes in each population. Corresponding RAPD phenotypes were obtained by determining the fragment lengths that would be obtained by RAPD analysis of each diploid individual. Presence of a fragment in either haplotype results in band presence in the diploid phenotype. Allele frequencies for each band were estimated from the observed phenotypes, and the nucleotide divergence was obtained from weighted counts of band sharing. The precision is notably less than that in the haploid case, especially when the sample size was 10 individuals, but, in the case of $s = 50$, the estimates of divergence were acceptable for small d (table 2 and fig. 3). In the case of a sample size of 10, the bias-corrected estimate of allele frequency improved the accuracy of the divergence estimates somewhat (data not shown). Estimates of the standard error of d obtained by jackknifing over the 10 primers were in good agreement with the standard error across replicates of the simulation. As in the haploid case, haplotypes are sampled from a neutral phylogeny, and all sites were considered as linked. Allowing recombination would reduce the variance among samples, in both the haploid case and the diploid case.

Table 1
Simulations of Pairs of Diverging Haploid Populations, and Estimates of Nucleotide Divergence *d*

Expected d[a]	Sample Size = 10[b]		Sample Size = 50[b]	
	d-seq[c]	d-RAPD[d]	d-seq[c]	d-RAPD[d]
0.01	0.017 ± 0.008	0.017 ± 0.009	0.014 ± 0.004	0.015 ± 0.005
0.02	0.025 ± 0.007	0.024 ± 0.008	0.024 ± 0.006	0.023 ± 0.005
0.03	0.035 ± 0.007	0.035 ± 0.010	0.036 ± 0.011	0.037 ± 0.014
0.04	0.041 ± 0.006	0.041 ± 0.008	0.042 ± 0.009	0.041 ± 0.011
0.05	0.055 ± 0.008	0.056 ± 0.010	0.058 ± 0.011	0.061 ± 0.016
0.06	0.066 ± 0.008	0.070 ± 0.014	0.065 ± 0.007	0.068 ± 0.013
0.07	0.074 ± 0.010	0.078 ± 0.019	0.072 ± 0.007	0.077 ± 0.020
0.08	0.079 ± 0.010	0.090 ± 0.023	0.082 ± 0.005	0.089 ± 0.015
0.09	0.089 ± 0.007	0.102 ± 0.025	0.091 ± 0.012	0.109 ± 0.032
0.10	0.096 ± 0.010	0.113 ± 0.026	0.097 ± 0.010	0.119 ± 0.027

[a] The theoretical expectation for nucleotide divergence.

[b] The no. of individuals in each population.

[c] The nucleotide divergence estimated from the sequences generated in the simulations.

[d] The nucleotide divergence estimated from the RAPD banding patterns predicted from the sequences, when the methods described in this paper are applied.

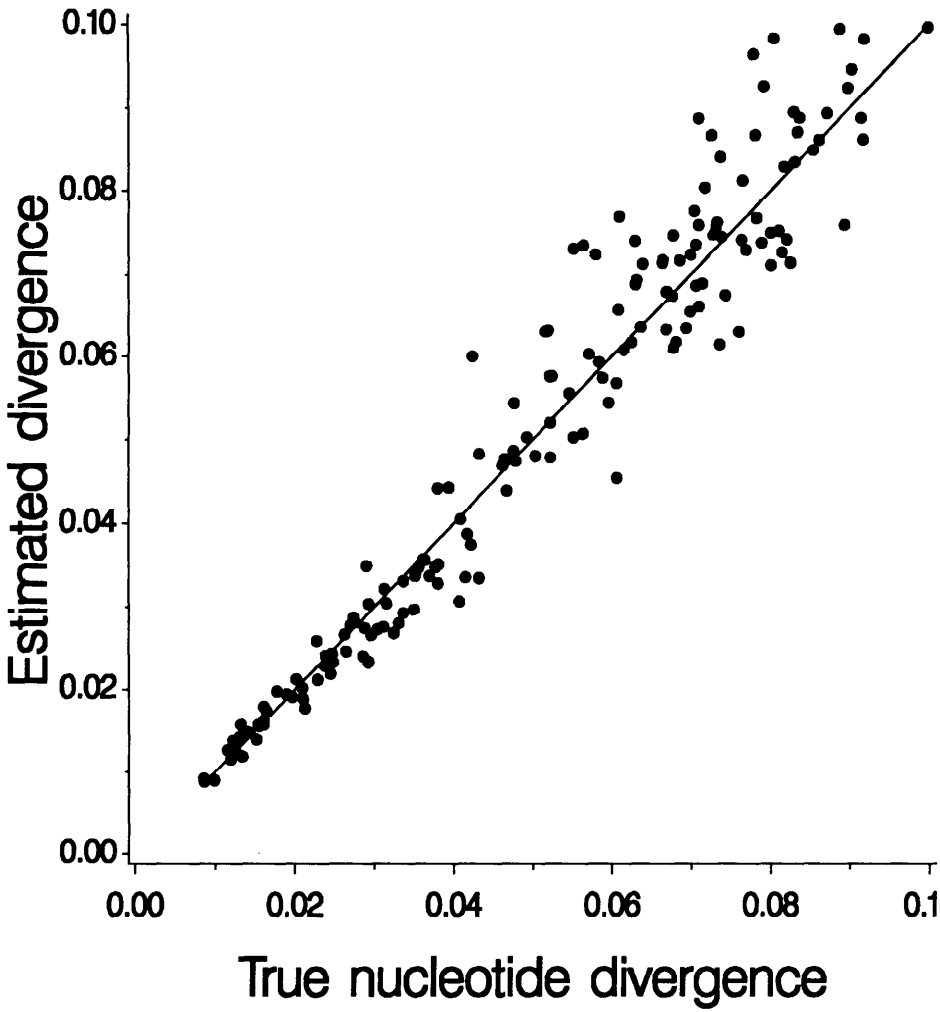


FIG. 2.—Simulations of nucleotide divergence in pairs of haploid populations. True divergence refers to the divergence in nucleotide sequences obtained from the simulations, and estimated divergence is based on the predicted RAPD banding patterns in the case with a sample of 50 alleles from each population. Details of the simulation are given in the text.

Insertions and Deletions

If insertions and deletions are frequent in a sample, then one can estimate the probability that two sequences mismatch because of insertion or deletion in a particular fragment. Tajima and Nei (1984) provide means of estimating this probability (which they call "γ") on the basis of restriction-fragment data, and these same approaches can be directly applied to RAPD data, provided that appropriate bands have been identified. Unfortunately, there is no good way to combine estimates of d and γ to provide a composite measure of nucleotide divergence. In species in which a substantial portion of polymorphism is due to insertions and deletions, such as *Drosophila melanogaster* (e.g., see Aquadro et al. 1986), RAPDs are likely to be unsuitable for estimating nucleotide divergence.

Table 2
Simulations of Pairs of Diverging Diploid Populations, and Estimates of Nucleotide Divergence *d*

Expected *d*	Sample Size = 10		Sample Size = 50	
	d-seq	*d*-RAPD	*d*-seq	*d*-RAPD
0.01	0.014 ± 0.008	0.011 ± 0.008	0.014 ± 0.004	0.016 ± 0.005
0.02	0.024 ± 0.008	0.019 ± 0.008	0.024 ± 0.006	0.024 ± 0.006
0.03	0.033 ± 0.007	0.029 ± 0.007	0.036 ± 0.011	0.037 ± 0.011
0.04	0.043 ± 0.005	0.039 ± 0.009	0.042 ± 0.009	0.041 ± 0.010
0.05	0.056 ± 0.008	0.051 ± 0.010	0.058 ± 0.011	0.059 ± 0.014
0.06	0.065 ± 0.010	0.063 ± 0.012	0.065 ± 0.007	0.064 ± 0.011
0.07	0.072 ± 0.008	0.074 ± 0.011	0.072 ± 0.007	0.072 ± 0.015
0.08	0.077 ± 0.007	0.078 ± 0.016	0.082 ± 0.005	0.081 ± 0.012
0.09	0.087 ± 0.012	0.089 ± 0.020	0.091 ± 0.012	0.096 ± 0.021
0.10	0.098 ± 0.009	0.099 ± 0.030	0.097 ± 0.010	0.105 ± 0.022

NOTE.—Definitions are as in table 1.

Conclusions

RAPDs may be useful for estimating nucleotide divergence of closely related taxa if a number of criteria are satisfied, including the following:

1. Primer selection must not be biased in favor of those that reveal the most polymorphism. Commercially available primers tend to have a G+C content of 60%–80%, which may result in an overestimate of human nucleotide diversity, because of the high degree of polymorphism at CpG sites.
2. All polymorphic and monomorphic bands must be carefully scored. If some bands are not scored, then there must be no bias in scoring monomorphic bands versus polymorphic bands.
3. Polymorphic bands must be shown to behave as Mendelian factors.
4. Allelism of bands must be ascertained by Southern blotting or segregation analysis. If two or more bands are allelic, only one should be scored.
5. Homology of bands of the same size in different species should be demonstrated, e.g., by Southern blotting.
6. For diploids, a population sample must be examined to determine band frequencies.
7. True nucleotide sequence divergence should not exceed ~10%.
8. Single nucleotide substitutions are assumed to result in a loss of amplification. We assume that amplification at imperfectly matching primer sites is rare, but further experimental work on both of these issues is desirable.
9. Insertion/deletion variation that results in variation in band presence/absence is assumed to be rare. Insertion/deletion variation that results in variation in band size must be identified and analyzed appropriately (see 4).

Even if one can identify bands that segregate as good Mendelian markers, DNA preparations of low quality may result in higher rates of mispriming, making it impossible to get an accurate count of monomorphic bands (Williams et al. 1990). For the purposes of estimation, the monomorphic bands should represent unvaried sites, which is not the case for PCR artifacts. Because the genetic identity of monomorphic

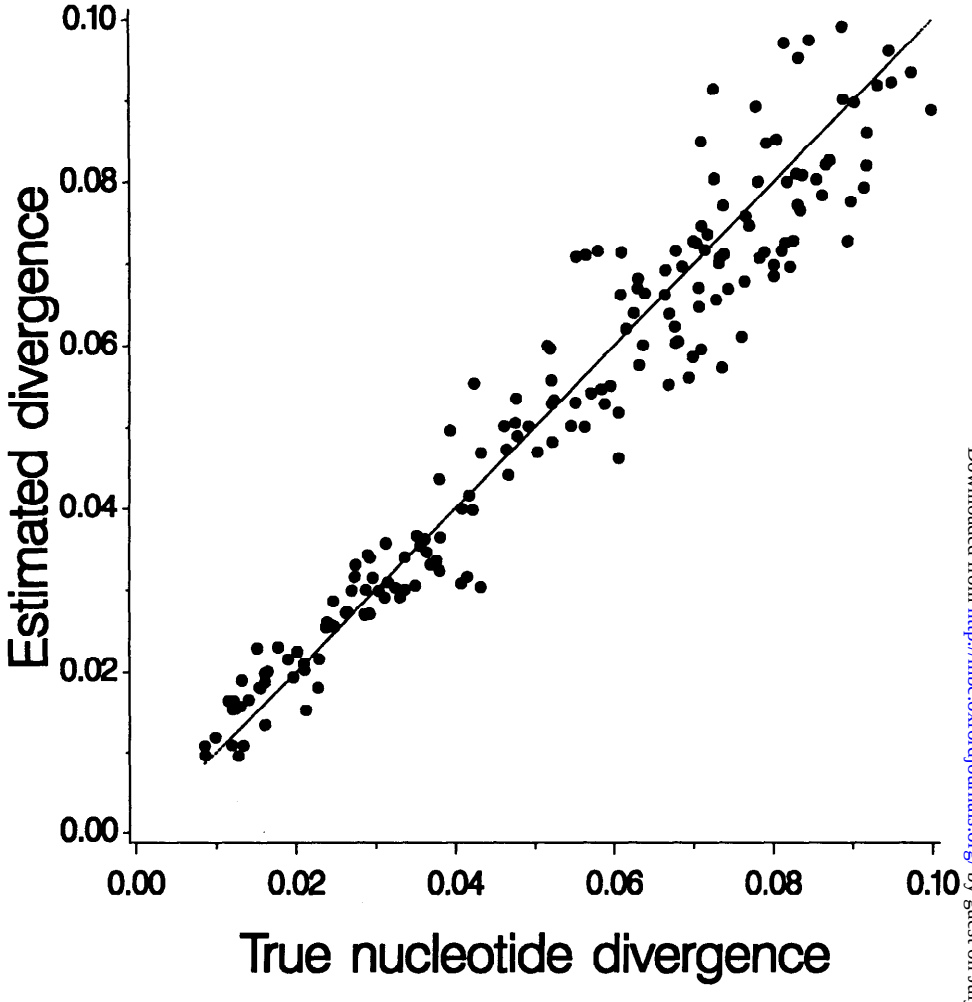


FIG. 3.—Simulations of nucleotide divergence for the diploid case. Estimated divergence values are from the case with a sample of 50 individuals from each population. Estimates of nucleotide divergence d from diploid data are somewhat less accurate than those from haploid data, in part owing to the dominance of RAPD band presence.

bands cannot be easily determined, high-quality DNA preparations and PCR reactions lacking non-Mendelian bands are essential. That estimates of d are accurate only when the true value is <0.10 limits application of the method to closely related taxa. Although RAPDs can be an efficient means of collecting large quantities of nucleotide divergence data, we emphasize that, unless these conditions are met, inference of phylogenetic relationships on the basis of RAPDs can be highly error prone.

Acknowledgments

We thank Drs. Masatoshi Nei, Tim Prout, Chuck Langley, Mike Lynch, Franjo Weissing, Odilia Velterop, and Tom Whittam for comments. Tatsuya Ota provided a computer program for generating nucleotide sequences sampled from a pair of diverging populations. Support included a Sloan Sabbatical award to A.G.C., an NSERC

grant to C.M.S.L., and NIH grants RR05090 and RR00169. This study was begun while A.G.C. was on sabbatical in Dr. Langley's laboratory at the University of California, Davis.

APPENDIX

The Expected Distribution of RAPD Bands, by Franz J. Weissing and Odilia Velterop, Department of Genetics, University of Groningen (Haren, The Netherlands)

We derive the expected distributions of the number and length of RAPD fragments generated from a random and unstructured DNA sequence by a primer that is neither self-overlapping nor palindromic. We assume that the probability a of matching a primer at a site x on a DNA strand is independent of the position of the site, the proximity of other matches, the orientation of the strand, and the occurrence of matches on the complementary strand. Let C denote the total number of nucleotide pairs of the double-stranded DNA sequence whose two strands will be called "plus" and "minus." Positions on the DNA sequence will be indicated by x, where $1 \leq x \leq C$. With respect to a fixed orientation, x refers to the $5' \rightarrow 3'$ direction on the plus strand and to the $3' \rightarrow 5'$ direction on the minus strand. A match of a given random primer with the plus or minus strand will be called a "plus match" or a "minus match," respectively. Neglecting the length of the primer, we assume that a RAPD fragment of length L is obtained whenever a minus match at position x is followed by a plus match at position $x + L$, with no other matches in between.

For a nonpalindromic primer, a plus match and a minus match cannot occur simultaneously at a given position. Accordingly, the probability that a given position has either a plus match or a minus match is $2a$. The probability that we obtain a fragment of length L nucleotides between a minus match and a plus match is

$$\text{Prob(fragment of length } L) = a^2(1 - 2a)^L . \tag{A1}$$

The probability that a RAPD fragment of any length is obtained is the sum of this quantity over all possible lengths, or

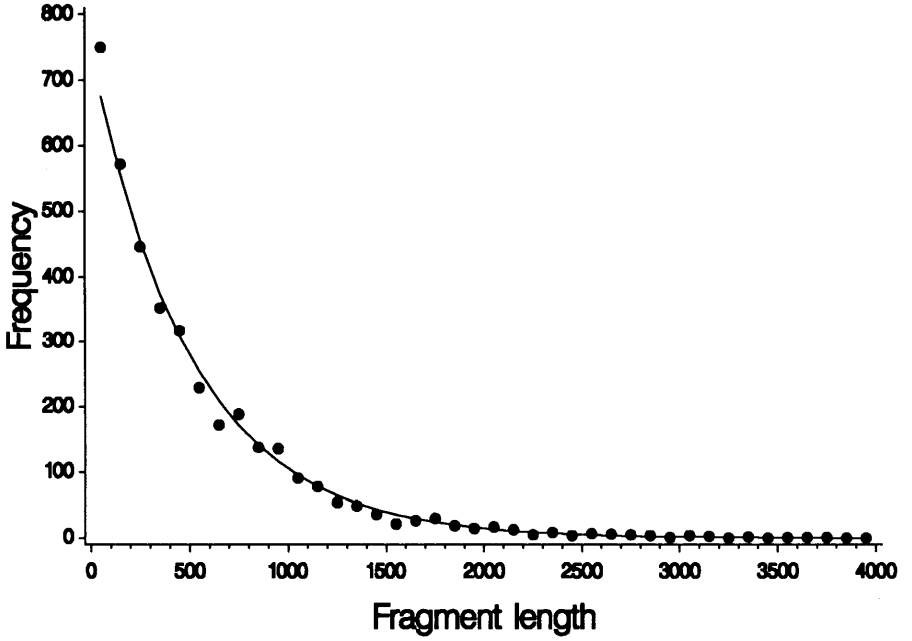
$$\text{Prob(fragment)} = \sum_L a^2(1 - 2a)^L = a(1 - 2a)/2 . \tag{A2}$$

The probability density function of RAPD fragment lengths (p_L) is therefore the probability that a fragment of length L is obtained, normalized by the probability of obtaining any fragment:

$$p_L = \frac{a^2(1 - 2a)^L}{a(1 - 2a)/2} = 2a(1 - 2a)^{L-1} \approx 2ae^{-2aL} . \tag{A3}$$

This exponential distribution has a mean fragment length of $1/(2a)$, and the variance in fragment length is $1/(2a)^2$. Fragment lengths obtained in simulation runs of the RAPD procedure fit well to this prediction (fig. A1).

To derive the distribution of the number of fragments, consider first the situation that there are exactly M matches. The probability of this event is given by



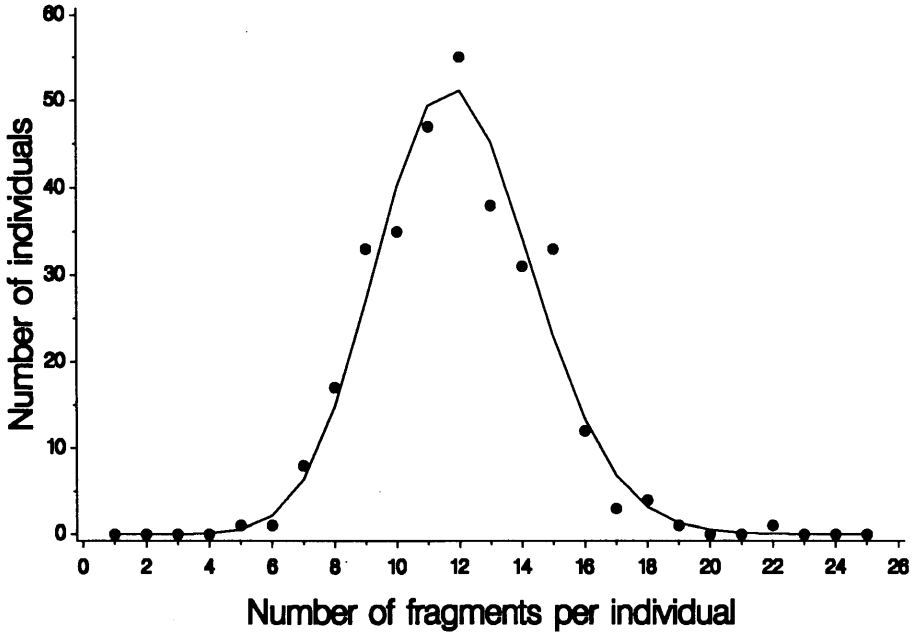


FIG. A1.—Comparison of the expected fragment length distribution (A3) and the expected fragment number distribution (A6) with the results of computer simulations of the RAPD procedure. For the simulations, 16 nonrepetitive, nonpalindromic primers of length 5 were combined with 20 randomly generated sequences of length $C = 25{,}000$, with all four nucleotide types equally frequent. A perfect match was required to obtain a fragment; thus $a = 1/4^5$. *Top,* Comparison of the length distribution of fragments formed in all 320 simulations with an exponential distribution of mean $1/(2a)$. According to distribution (A9), $aC/2 = 12.2$ fragments are expected per simulation with a variance of $aC/4 = 6.1$. *Bottom,* Fragment number

$$\text{Prob}(M \text{ matches}) = e^{-2aC}\frac{(2aC)^M}{M!}. \tag{A4}$$

Since matches are symbolized by a plus sign or a minus sign, the pattern of matches can be characterized by a sequence of M signs. The number of RAPD fragments is fully determined by this sign sequence, because a fragment is generated whenever a minus sign is followed by a plus sign. Let us therefore leave the context of DNA sequence of length C and focus on a sign sequence of length M. It is useful to imagine a sign sequence as consisting of clusters of plus signs separated by clusters of minus signs. The sequence is fully determined by the positions at which the transitions from plus to minus and from minus to plus occur. Notice that the number of minus-to-plus transitions corresponds to the number of RAPD fragments. We will show next that the probability that a sign sequence of length M contains exactly m minus-to-plus transitions is given by

$$\text{Prob}(m \text{ fragments} \mid M \text{ matches}) = \frac{1}{2^M}\binom{M+1}{2m+1}. \tag{A5}$$

The first factor in probability (A5) is explained by the fact that there are 2^M possible sign sequences of length M. To explain the second factor, we consider a prolonged sign sequence where a plus sign is added at position 0 and a minus sign is added at position $M + 1$. The prolonged sign sequence always starts and ends with a plus-to-minus transition, whereas the number of minus-to-plus transitions (and thus the number of fragments generated) remains unchanged. Accordingly, there are m minus-to-plus and $m + 1$ plus-to-minus transitions that may occur at the positions $1, 2, \ldots, M + 1$ of the prolonged sign sequence. The second factor in probability (A5) corresponds to the number of ways in which $2m + 1$ sign transitions can be distributed over $M + 1$ positions.

The probability q_m that m fragments are generated is obtained by combining probabilities (A4) and (A5):

$$q_m = \sum_M e^{-2aC}\frac{(2aC)^M}{M!}\frac{1}{2^M}\binom{M+1}{2m+1} = e^{-aC}\left(\frac{(aC)^{2m}}{(2m)!} + \frac{(aC)^{2m+1}}{(2m+1)!}\right). \tag{A6}$$

The mean and variance of this distribution are given by

$$\mu = \frac{aC}{2} - \frac{1}{4}(1 - e^{-2aC}) \tag{A7}$$

and

$$\sigma^2 = \frac{aC}{4}(1 - 2e^{-2aC}) + \frac{1}{16}(1 - e^{-4aC}). \tag{A8}$$

Thus, to a good approximation, the mean and variance of the distribution of fragment number are related to the matching probability a and the sequence length C by

$$\mu \approx \frac{aC}{2}, \qquad \sigma^2 \approx \frac{aC}{4} = \frac{\mu}{2}. \tag{A9}$$

Notice that the variance of the fragment number distribution is only half as large as the variance of a Poisson distribution with mean $(aC)/2$. Figure A1 shows that the number of fragments generated by computer simulations of the RAPD procedure do fit well to the distribution given by probability (A5).

LITERATURE CITED

AQUADRO, C. F., S. F. DESSE, M. M. BLAND, C. H. LANGLEY, and C. C. LAURIE-AHLBERG. 1986. Molecular population genetics of the *alcohol dehydrogenase* gene region of *Drosophila melanogaster*. Genetics **114**:1165–1190.

BERNARDI, G. 1989. The isochore organization of the human genome. Annu. Rev. Genet. **23**: 637–661.

BERNARDI, G., B. OLOFSSON, J. FILIPSKI, M. ZERIAL, J. SALINAS, G. CUNY, M. MEUNIER-ROTIVAL, and F. RODIER. 1985. The mosaic genome of warm-blooded vertebrates. Science **228**:953–958.

BULMER, M. 1986. Neighboring base effects on substitution rates in pseudogenes. Mol. Biol. Evol. **3**:322–329.

EFRON, B. 1982. The jackknife, the bootstrap, and other resampling plans. Society of Industrial and Applied Mathematics, Philadelphia.

FELLER, W. 1968. An introduction to probability theory and its applications, 3d ed. Vol. **2**. Wiley, New York.

HADRYS, H., M. BALICK, and B. SCHIERWATER. 1992. Applications of random amplified polymorphic DNA (RAPD) in molecular ecology. Mol. Ecol. **1**:55–63.

LANIGAN, C. M. S. 1992. RAPD analysis of primates: phylogenetic and genealogical considerations. Ph.D. diss., Genetics Graduate Group, University of California, Davis.

LYNCH, M. 1990. The similarity index and DNA fingerprinting. Mol. Biol. Evol. **7**:478–484.

LYNCH, M., and B. G. MILLIGAN. Analysis of population genetic structure with RAPD markers (submitted).

MARTIN, G. B., J. G. K. WILLIAMS, and S. D. TANKSLEY. 1991. Rapid identification of markers linked to a *Pseudomonas* resistance gene in tomato by using random primers and near isogenic lines. Proc. Natl. Acad. Sci. USA **88**:2336–2340.

NEI, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York.

NEI, M., and W.-H. LI. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA **76**:5269–5273.

NEI, M., and J. C. MILLER. 1990. A simple method for estimating average number of nucleotide substitutions within and between populations from restriction data. Genetics **125**:873–879.

NEI, M., and F. TAJIMA. 1983. Maximum likelihood estimation of the number of nucleotide substitutions from restriction sites data. Genetics **105**:207–217.

RIEDY, M. F., W. J. HAMILTON, and C. F. AQUADRO. 1992. Excess of non-parental bands in offspring from known primate pedigrees assayed using RAPD PCR. Nucleic Acids Res. **20**: 918.

TAJIMA, F., and M. NEI. 1984. Estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. **1**:269–285.

TAKAHATA, N., and M. NEI. 1985. Gene genealogy and variance of interpopulational nucleotide differences. Genetics **110**:325–344.

WATERMAN, M. S. 1983. Frequencies of restriction sites. Nucleic Acids Res. **11**:8951–8956.

WELSH, J., and M. MCCLELLAND. 1990. Fingerprinting genomes using PCR with arbitrary primers. Nucleic Acids Res. **18**:7213–7218.

WILLIAMS, J. G. K., A. R. KUBELIK, K. J. LIVAK, J. A. RAFALSKI, and S. V. TINGEY. 1990.

DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res. **18**:6531–6535.

WILLIAMS, J. G. K., M. K. Hanafey, J. A. RAFALSKI, and S. V. TINGEY. 1993. Genetic analysis using random amplified polymorphic DNA markers. Methods Enzymol. **218**:704–740.

MASATOSHI NEI, reviewing editor

Received April 27, 1992; revision received February 23, 1993

Accepted February 25, 1993