DNA Fingerprinting: Approaches and Applications
ed. by T. Burke, G. Dolf, A. J. Jeffreys & R. Wolff

Analysis of Population Genetic Structure by DNA Fingerprinting

M. Lynch

Department of Biology, University of Oregon, Eugene, Oregon 97403, USA

Summary
DNA fingerprint similarity is now being used widely to make inferences about the genetic structure of natural and domesticated populations, often with little regard to the limitations of such data. This paper provides an overview of the statistical theory of DNA fingerprint analysis with special focus on applications to natural populations for which little if anything is known about the detailed genetics of the DNA profiles. Approaches to estimating individual and population homozygosity, effective population size, population subdivision, and relatedness are reviewed, and issues concerning the biases and sampling properties of the statistics are discussed.

Introduction

Tests of a number of ideas in evolutionary biology, particularly in areas of social organization and kin selection, require accurate estimates of relatedness between individuals, of levels of individual and population homozygosity, and of the degree of population subdivision. Similar types of measurements are needed in programs of genetic conservation, as are estimates of pedigree structure and of breed differentiation. Traditionally, isozymes and blood proteins have been exploited for these purposes, their advantage being a clear Mendelian interpretation of banding patterns on gels. However, protein markers also have significant disadvantages, most notably the need for different protocols for each locus, relatively low levels of detectable polymorphisms, and the weak statistical power associated with loci exhibiting low degrees of variation.

Thus, it comes as no surprise that hypervariable DNA-fingerprinting loci (Jeffreys *et al.*, 1985b, c; Jeffreys *et al.*, 1990) have been embraced widely as a sort of mother lode of genetic information for studies on the structure of natural and domesticated populations. Because such loci usually exist as dispersed families with a common core sequence, multiple restriction fragment length polymorphisms can be visualized simultaneously on the same gel. This has substantial economic advantages. Moreover, the high levels of allelic diversity at DNA-fingerprinting loci imply a maximum amount of information obtained per unit effort. However, one pays a price for these apparent advantages. Although

multilocus profiles have a Mendelian basis, an exact genetic interpretation is usually beyond reach. Without laborious breeding experiments, specific bands cannot be associated with particular loci, and fragments with low molecular weights will usually go undetected. Consequently, locus specific gene and genotype frequencies, the basis for all conventional population genetic analyses, are usually unknown. Alternative analytical procedures are required for DNA-fingerprinting studies.

This paper provides an overview of the statistical issues associated with DNA-fingerprint analysis, some of which are covered in more technical detail in Lynch (1988, 1990), Brookfield (1989), and Cohen (1990). Special attention will be given to the types of inferences that can be made in the absence of information or assumptions on allele-frequency distributions, since this will usually be a necessity for the practitioner. There are many aspects of sample preparation, gel running and reading that can lead to problems before the analysis of data even begins (Lander, 1989), but these are ignored below in order to focus on the essential mathematical issues.

The DNA-Fingerprint Phenotype

The fundamental units of data in a DNA-fingerprinting study are the numbers of bands exhibited in individual lanes n_x, where x denotes an individual, and the number of shared bands for pairs of individuals n_{xy}. Although most applications of DNA-fingerprinting have focused on aspects of band sharing, the average number of bands can also provide useful information about population structure. If L is the average number of loci sampled, then the expected number of bands for individual x is

$$E(n_x) = L(2 - H_x), \tag{1}$$

where H_x is the homozygosity of individual x at an average fingerprinting locus. This formula also applies to the average number of bands for random members of a population when the subscript x is dropped and H is taken to be the mean homozygosity in the population. Since homozygosity is related to the level of inbreeding, a simple investigation of the average number of bands in fingerprint profiles may be a useful means of assessing variation in inbreeding within and between populations.

It seems reasonable to assume that the length variants at VNTR loci are effectively neutral, in which case Malécot's (1948) recursion equation

$$H_t = (1 - \mu)^2\left[\frac{1}{2N} - \left(1 - \frac{1}{2N}\right)H_{t-1}\right] \tag{2}$$

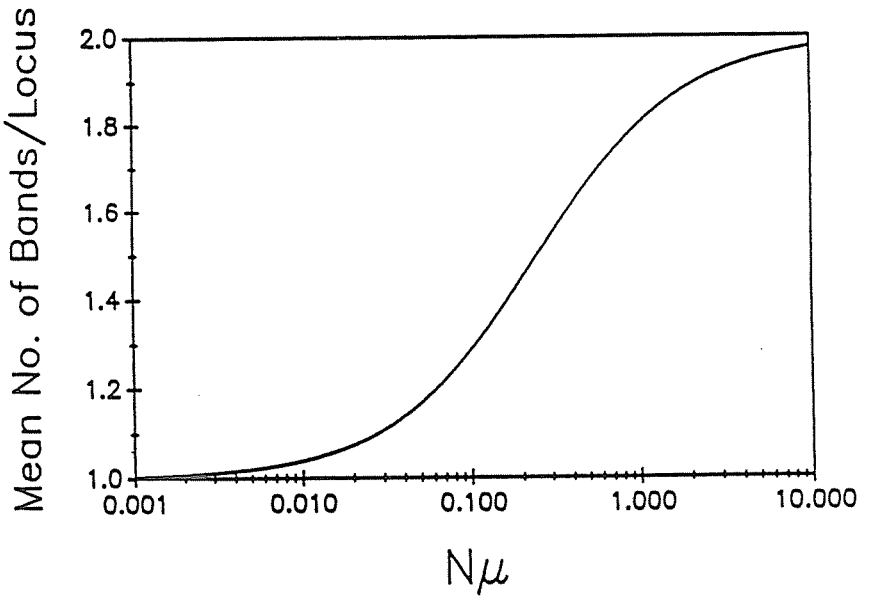

Figure 1. Relationship between the mean number of bands per locus for individuals from a population in drift-mutation equilibrium and the composite parameter $N\mu$. N is the effective population size and μ is the gametic mutation rate.

where N is the effective population size and μ is the mutation rate (to a new discernible length variant), can be used to project the dynamics of homozygosity under the joint influence of mutation and random genetic drift. Under drift-mutation equilibrium, $\hat{H} \simeq 1/(1+4N\mu)$, and the expected number of bands per individual is

$$E(\hat{n}) = L\left(\frac{8N\mu + 1}{4N\mu + 1}\right). \tag{3}$$

Figure 1 shows that the average number of bands is a most sensitive indicator of the composite parameter $N\mu$ when the latter is in the range of 0.1 to 1.0.

Rearrangement of Equation (3) yields an estimator for the effective size of an equilibrium population in terms of the average number of bands per individual,

$$N = \frac{\bar{n} - L}{4\mu(2L - \bar{n})}. \tag{4}$$

Application of this formula requires an estimate of the mutation rate, which can in principle be obtained by observing the incidence of nonparental bands in progeny (Jeffreys *et al.*, 1985c, 1988; Gyllensten *et al.*, 1989; Georges *et al.*, 1990). Direct estimation of the number of loci sampled is more difficult in the absence of breeding experiments, but to

a first approximation $L \simeq \bar{n}(4-\bar{S})/4(2-\bar{S})$, where $\bar{S}$ is the average fraction of shared bands for pairs of nonrelatives (Lynch, 1990). Noting that this expression tends to slightly underestimate L and substituting into Equation (4),

$$N = \frac{4-3\bar{S}}{8\mu\bar{S}} \tag{5}$$

provides an upwardly biased estimate for the effective population size.

It is also possible to use Equation (2) to project the change in the mean number of bands per individual as a population becomes progressively inbred. Letting H_0 be the homozygosity at time zero, $\lambda = 1 - (1/2N)$, and $\phi = \lambda(1-\mu)^2$,

$$H_t = \frac{(1-\mu)^2}{2N} + H_0\phi^t + \frac{1}{2N-1}\sum_{i=2}^{t}\phi^i. \tag{6}$$

Figure 2 shows the response of the mean number of bands to prolonged periods of small effective size as a function of the conventional inbreeding coefficient $f_t = 1 - \lambda^t$. Note that the expected number of bands declines with increasing f in a roughly linear manner. However, $\bar{n}$ is not

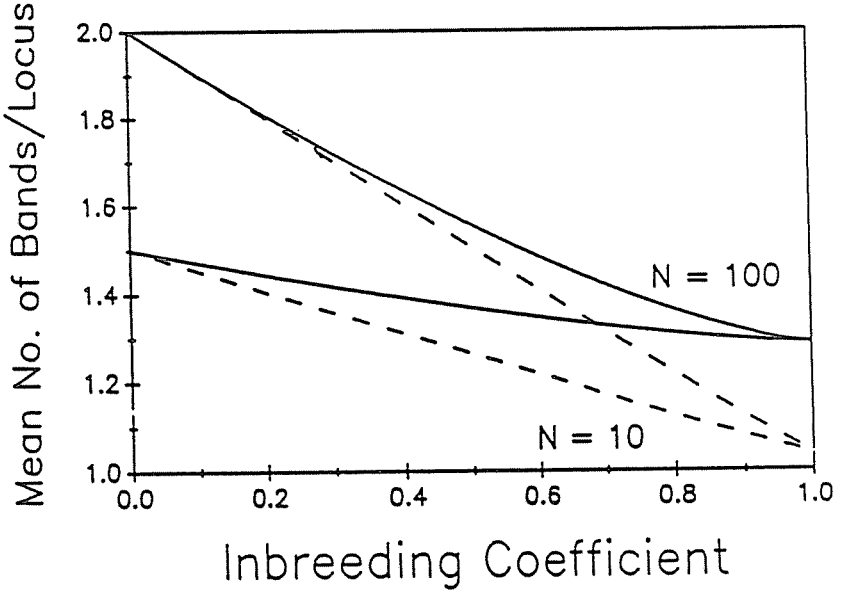

Figure 2. The transient response of the mean number of bands per locus to prolonged periods at effective population sizes of 10 (dashed lines) and 100 (solid lines). The upper and lower sets of lines refer to situations in which the base population is 100% and 50% heterozygous respectively. As the inbreeding coefficient approaches one, a new equilibrium mean number of bands per locus is established, as defined by Equation (3). The inbreeding coefficient can be viewed as the expected homozygosity at more stable loci that have a negligible chance of mutating over periods of time $< t/N$ generations.

just a function of f. Larger populations have higher values of $\bar{n}$ than smaller populations with the same inbreeding coefficient due to the greater opportunities for the replenishment of alleles by mutation in the former. Note also that a substantial amount of heterozygosity can exist at VNTR loci in an equilibrium population, due to high mutation rates, while more stable loci are essentially 100% inbred (and homozygous, as denoted by f).

The Similarity Index

The similarity index for individuals x and y is the number of common bands in their fingerprint profiles divided by the average number of bands exhibited by both individuals,

$$S_{xy} = \frac{2n_{xy}}{n_x + n_y}. \tag{7}$$

For randomly mating population in genetic equilibrium within and between loci, the expected similarity for random pairs of individuals is

$$E(S) = \frac{\sum_{k,i} p_{ki}^2(2 - p_{ki})}{L} \tag{8}$$

where p_{ki} is the frequency of the ith allele at the kth locus (Jeffreys *et al.*, 1985a; Lynch, 1988). This shows that the average similarity does not have a conventional interpretation from the standpoint of population genetics. Since the term $(2 - p_{ki})$ is always greater than one, the mean similarity always overestimates the population homozygosity, $\Sigma_{k,i} p_{ki}^2/L$, with the inflation being approximately two-fold when all alleles are rare. Usually, all alleles are not rare, in which case the bias is not so great, but in any event its magnitude cannot be determined in the absence of information on allele frequencies.

On the other hand, over a broad spectrum of gene frequency distributions, the mean similarity does approximate the average identity-in-state-of pairs of individuals (Lynch, 1990). For any locus, identity-in-state of pairs of individuals is either, 0, 0.5, or 1.0, depending on whether the genotypes share 0, 1, or 2 of the same genes. For random members of a panmictic population,

$$E(I) = E(S) - \frac{\sum_{k,i} p_{ki}^3(1 - p_{ki})}{L} \tag{9}$$

Thus, as in the case of population homozygosity, the similarity index is always an upwardly biased estimator of I. However, the bias is a function of the cubed gene frequencies, attaining a maximum of

$E(S) - E(I) = 0.125$ when $p = 0.5$ for all alleles, which is substantially less than $E(S) - E(H) = 0.25$ which arises under the same conditions.

Recalling that the mean similarity somewhat overestimates the population homozygosity and substituting for the expected value of the latter under drift-mutation equilibrium,

$$N = \frac{1 - \bar{S}}{4\mu S} \tag{10}$$

provides a downwardly biased estimate of the effective population size. An advantage of this expression over Equation (5) is that its derivation does not require an assumption regarding the number of loci sampled. Together, the two formulae should provide an order-of-magnitude approximation of the effective population size, provided the assumption of drift-mutation equilibrium is met.

Sampling Variance of the Basic Statistics

DNA-fingerprint profiles usually involve fairly large numbers (20 to 40) of individually segregating fragments. Thus, by the central limit theorem, the composite statistics n_x and S_{xy} should be roughly normal in distribution, in which case standard statistical procedures can be used to construct confidence limits for the population means of these quantities.

For example, assuming samples have been taken randomly, the sampling variance of the number of bands per individual is simply

$$\mathrm{Var}(n) = \frac{k(\overline{n^2} - \bar{n}^2)}{k - 1}, \tag{11}$$

where k is the sample size. The sampling variance of the average number of bands is this quantity divided by k. The standard errors of n and $\bar{n}$, which are the square roots of the sampling variances, can be used to construct confidence limits and for other applications associated with hypothesis testing. For instance, if the assumption of drift-mutation equilibrium is valid, populations that differ significantly in $\bar{n}$, must also differ in $N\mu$, most likely in N. The confidence limits for n and $\bar{n}$ are obtained by multiplying the square roots of the sampling variances by the appropriate t values (obtained from any basic statistics text).

Depending on what individuals are used to estimate similarities, computation of the sampling variance of $\bar{S}$ is somewhat more involved. If individuals are used in multiple comparisons (as when all possible combinations of lanes on a gel are compared), the data will not be independent. Similarity measures involving a common member tend to be positively correlated–individuals that happen to exhibit bands that are relatively common in the population will tend to have high similarities with most other individuals, and *vice versa* for those that happen to

carry rare alleles. The standard formula for a sampling variance assumes independence of data, and in this case, its application would yield downwardly biased estimates.

To cope with this problem, the investigator has two options. On the one hand, each individual similarity estimate could be based on a unique, nonoverlapping pair of individuals (i.e., 1 and 2, 3 and 4, etc.), and the standard variance formula used. In that case, the total number of similarity estimates can be no greater than half the number of individuals sampled, a sacrifice that most investigators will not want to make. Alternatively similarities can be estimated for arbitrary pairs of individuals, and the sampling variance computed by the formula of Lynch (1990),

$$\mathrm{Var}(\bar{S}) = \frac{k\mathrm{Var}(S_{xy}) + 2k'\mathrm{Cov}(S_{xy}, S_{xz})}{k^2} \tag{12}$$

where k is the total number of similarity measures used to estimate $\bar{S}$ and k' is the number of pairs of those measures that share an individual. For example, if all possible comparisons between four individuals have been made $k = 6$ and $k' = 12$. The standard error of $\bar{S}$ is estimated by the square root of this quantity.

In Equation (12), $\mathrm{Var}(S_{xy})$ is the unbiased estimate of the variance of independent similarity measures. It can be estimated with

$$\mathrm{Var}(S_{xy}) = \frac{\sum(S_{wx} - S_{yz})^2}{2k^*}. \tag{13}$$

where k^* is the number of pairwise comparisons that do not share members. This is a more general formula than that given in Lynch (1990) in that it uses all applicable pairs of data. The sampling covariance of overlapping similarities can also be estimated directly from the data,

$$\mathrm{Cov}(S_{xy}, S_{xz}) = \frac{k^*(\overline{S_{xy}S_{xz}} - \bar{S}^2)}{k^* - 1}, \tag{14}$$

where k^* is now the number of pairs of comparisons involving shared members. The mean cross-product can be computed most efficiently by focusing on adjacent triplets on gels (i.e., lanes 1, 2 and 3 yield $S_{12}S_{23}$, lanes 4, 5 and 6 yield $S_{45}S_{56}$, etc.) The estimate of $\bar{S}$ to be used in this formula should be based on the same measures as the mean cross-product.

It should be noted that the Equation (12) only estimates the sampling variance of $\bar{S}$ associated with the loci that happened to be included in the survey. It does not account for the error arising from the sampling of a finite number of loci. If, however, one is willing to assume that the sampled loci have gene-frequency distributions that are representative

of other such loci throughout the genome, then

$$\mathrm{Var}'(S_{xy}) = \frac{2\bar{S}(1-\bar{S})(2-\bar{S})}{\bar{n}(4-\bar{S})} \tag{15}$$

accounts for this additional source of sampling error (Lynch, 1990). Equation (15) should be used in place of Equation (13) when the mean similarities of different populations are being compared and it is uncertain whether the same loci have been sampled. It should also be used when one is using the set of sampled loci to make inferences about genome wide properties. Since the covariance between similarity measures is proportional to the sampling variance, Equation (14) can be corrected for locus sampling by multiplying by $\mathrm{Var}'(S_{xy})/\mathrm{Var}(S_{xy})$. Although Equation (15) is only a first-order approximation, it tends to slightly overestimate the actual sampling variance and therefore yields conservative estimates for the standard error (Lynch, 1990).

Finally, it should be noted that problems of gel running and reading will ordinarily lead to some sampling variance in the estimates of S_{xy}, particularly when individuals from distant lanes and/or different gels are compared. Such variance has little to do with the genetic properties of the population. Thus, if one has an interest in the true genetic variation of n and/or S, the variance due to technical problems needs to be subtracted from the estimates described above. The latter variance can be estimated by scoring replicate pairs of individuals and computing the variance among replicates.

Population Subdivision

Measures of gene diversity (Nei, 1987) extracted from single-locus analyses are used frequently to estimate measures of population subdivision analogous to Wright's (1951) F-statistics. Strictly speaking, the usual formulae cannot be applied to DNA-fingerprinting data since explicit estimates of gene frequencies are not usually available. Nevertheless, it is possible to test the hypothesis of population subdivision through the use of the similarity index. A question of interest is whether there is significantly less similarity between samples from two populations than expected on the basis of the within-sample similarity. This can be resolved by computing an index of between-population similarity corrected by the within-population similarity,

$$\bar{S}_{ij} = 1 + \bar{S}'_{ij} - \frac{\bar{S}_i + \bar{S}_j}{2}, \tag{16}$$

where $\bar{S}_i$ is the average similarity of individuals within population i, and $\bar{S}'_{ij}$ is the average similarity between random pairs of individuals across populations i and j (Lynch, 1990). When $\bar{S}'_{ij}$ equals the mean similarity

in the two populations, $\bar{S}_{ij} = 1$ indicating that the populations are homogeneous. Lynch (1990) presents procedures for computing the sampling variance of $\bar{S}_{ij}$, which are necessary for testing the hypothesis of population subdivision. Problems of nonindependence, mentioned above, need to be accounted for in these computations.

As noted above, the similarity index yields upwardly biased estimates of population homozygosity. Hence, $1 - \bar{S}$ is a downwardly biased estimate of the heterozygosity (or gene diversity). Wright's (1951) index of population subdivision, F_{ST}, is defined to be the fraction of total gene diversity that is attributable to population differentiation. Since the similarity index will bias the estimates of both the within- and between-population homozygosity in the same direction, it seems likely that these biases will nearly cancel out when ratios of the components are employed. Thus, as a first-order approximation,

$$F'_{ST} \simeq \frac{1 - S_b}{2 - S_w - S_b}, \tag{17}$$

where S_b is the average valus of $\bar{S}_{ij}$ over all pairs of populations i, j, and S_w is the average value of $\bar{S}_i$ over all i. F'_{ST} takes on a maximum value of one when populations are fixed for different alleles and a minimum value of zero when there is no subdivision. A standard error for F'_{ST} can be obtained by use of a Taylor expansion approximation that takes into account the sampling variance-covariance structure of S_b and S_w (Lynch, 1990).

Two simple examples show that Equation (17) may work quite well under a broad range of conditions. Suppose that the true mean within-population homozygosity is 0.6 whereas the estimate is $S_w = 0.7$, and the true between-population homozygosity is 0.2 whereas the estimate $S_b = 0.4$. In that case, both the true value of F_{ST} and F'_{ST} are equal to 2/3. If, on the other hand, the between-population homozygosity is 0.8 and $S_b = 0.9$, $F'_{ST} = 1/4$ whereas the actual subdivision is $F_{ST} = 1/3$. In practice, differences of this magnitude are usually well within the bounds of sampling error.

F_{ST} is a measure of inbreeding due to population subdivision. Inbreeding may also result from consanguineous matings within populations. This is usually quantified by Wright's (1951) F_{IS}, which measures the fractional loss of heterozygosity due to local inbreeding on a scale of zero to one. For inbred populations, Equations (1) and (8) generalize to

$$E(n_F) = (1 - F_{IS})E(n_0) + LF_{IS}, \tag{18}$$

$$E(S_F) = E(S_0) + F_{IS} \sum_{k,i} p_{k.i}^2 (p_{k.i} - 1)/L. \tag{19}$$

These formula show that there is no simple way to quantify the degree of local inbreeding from direct observations of the average number of

bands per individual or of average similarity. If, however, a sample of the population can be mated randomly, then F_{IS} can be estimated from the observed mean number of bands before ($\bar{n}_F$) and after ($\bar{n}_0$) mating,

$$F_{IS} \simeq \frac{\bar{n}_F - \bar{n}_0}{L - \bar{n}_0}, \tag{20}$$

provided an estimate of the number of loci is available (see above). In the next section, a simple method for testing for consanguineous mating, which does not require an artificial breeding program, is introduced.

For the same reason that the similarity index can be used to obtain a nearly unbiased estimate of F_{ST}, it should also be possible to closely approximate the genetic distance between populations. An analog of Nei's (1972) estimator is

$$D'_{ij} \simeq -\ln\left(\frac{\bar{S}'_{ij}}{\sqrt{\bar{S}_i \bar{S}_j}}\right). \tag{21}$$

An expression for the sampling variance of D'_{ij}, which requires the use of some formulae in Lynch (1990), can be found in Nei (1987).

Under the assumption of drift-mutation-migration equilibrium, relatively simple expressions exist for expected values of F_{ST} and D in terms of migration and mutation rates (Nei, 1987), so Equations (17) and (21) may be of some use in estimating these parameters. For example, for populations that have been completely isolated for t generations, the expected value of D is $2\mu t$.

Estimation of Relatedness

The fact that most individuals in outbred populations have unique DNA-fingerprint profiles has encouraged the belief that the enormous power for identifying individuals would extend to identifying specific kinds of relationships. For a few types of applications, such as identification and/or exclusion of parentage, DNA fingerprinting has, in fact, been highly profitable (Jeffreys *et al.*, 1985a; Wetton *et al.*, 1987; Brookfield, 1989; Burke *et al.*, 1989; Morton *et al.*, 1990). However, successful extensions to more distant relationships are still notably absent.

It stands to reason that the DNA-fingerprint similarity between a pair of individuals should increase with their degree of relatedness. Lynch (1988) showed formally that this relationship is linear in a randomly mating population,

$$E(S) = \bar{\theta} + r(1 - \bar{\theta}), \tag{22}$$

where the relatedness r is the proportion of genes identical by descent between two individuals ($r = 0.5$ for parent-offspring, 0.25 for grand-

parent-grandchild and half-sibs, etc.), and $\bar{\theta}$ is the fraction of bands shared by nonrelatives. This formula states quite simply that the expected similarity of a pair of individuals is the sum of the probability that genes in the two individuals are identical by descent and the probability that they are not identical by descent but nevertheless identical in state. Equation (22) shows that a regression of S or r should have an intercept equal to $\bar{\theta}$ and a slope equal to $(1-\bar{\theta})$. Thus, provided the population is panmictic, the regression of S on r, which can be exploited in future studies, does not actually require the availability of pairs of individuals of various known degrees of relatedness. It merely requires an accurate estimate of the mean similarity between nonrelatives, a quantity that should ordinarily be obtainable.

In principle, S would provide a nearly unbiased estimate of r if the investigator were able to exploit a set of loci for which $\bar{\theta}$ is nearly zero. When most loci contain a large number of alleles, the genes in two individuals are unlikely to be identical in state unless they are also identical by descent. Unfortunately, for most existing studies $\bar{\theta}$ is on the order of 0.2 or much greater (Table 1), in which case the similarity index substantially exceeds the relatedness. Moreover, the magnitude of the bias increases with the more distant relationships–the very ones that it had been hoped that DNA fingerprinting would elucidate.

At first sight, one might expect that the bias could be eliminated by inserting the population estimate of $\bar{\theta}$ into Equation (22) and solving for r as a function of S. This should indeed work on average. There are, however, a couple of additional problems if one desires to estimate the relationships between specific pairs of individuals. First, the bias between S_{xy} and r_{xy} is not simply a function of the population average $\bar{\theta}$ but of the fraction of bands that the specific individuals x and y share

Table 1. Average fraction of shared bands for pairs of nonrelatives

Natural populations:		
House sparrows	0.1–0.3	Burke and Bruford 1987
Pied flycatchers	0.2	Wetton *et al.* 1987
Dunnocks	0.2	Burke *et al.* 1989
Purple martins	0.2	Morton *et al.* 1990
Channel Island foxes	0.7–1.0	Gilbert *et al.* 1990
Humans	0.2	Jeffreys *et al.* 1985c
Domesticated species:		
Chickens	0.4–1.0	Kuhnlein *et al.* 1990
Dogs	0.5	Jeffreys and Morton 1986
Cats	0.5	
Cattle	0.3–0.4	Georges *et al.* 1988
Horses	0.3–0.7	
Pigs	0.5–0.7	

with nonrelatives,

$$r_{xy} = \frac{E(S_{XY} - \theta_{xy})}{1 - \theta_{xy}} \tag{23}$$

(Lynch, 1988). θ_{xy} may be greater or less than $\bar{\theta}$ depending upon whether x and/or y carry relatively common or rare alleles. θ_{xy} will often be an unobservable quantity, and even when estimatable, will necessarily be less accurate than $\bar{\theta}$. Second, the range of expected similarity values is $\bar{\theta}$ to 1. Consequently, as the average similarity between nonrelatives increases, the regression of S on r becomes shallower, i.e., S becomes a less sensitive indicator of r. This is a serious issue since the variance among similarity values for specific kinds of relationships is quite high, even for parents and offspring if $\bar{\theta}$ is moderately high (Lynch, 1988).

For studies in which there is a moderate amount of similarity between nonrelatives (as is the case in almost all existing studies), there is little question that the distribution of similarity measures from adjacent types of relatives (e.g., parent-offspring vs. grandparent-grandchild) will be broadly overlapping. Assuming that the distribution of similarity is approximately normal, one can use Equation (22) in conjunction with

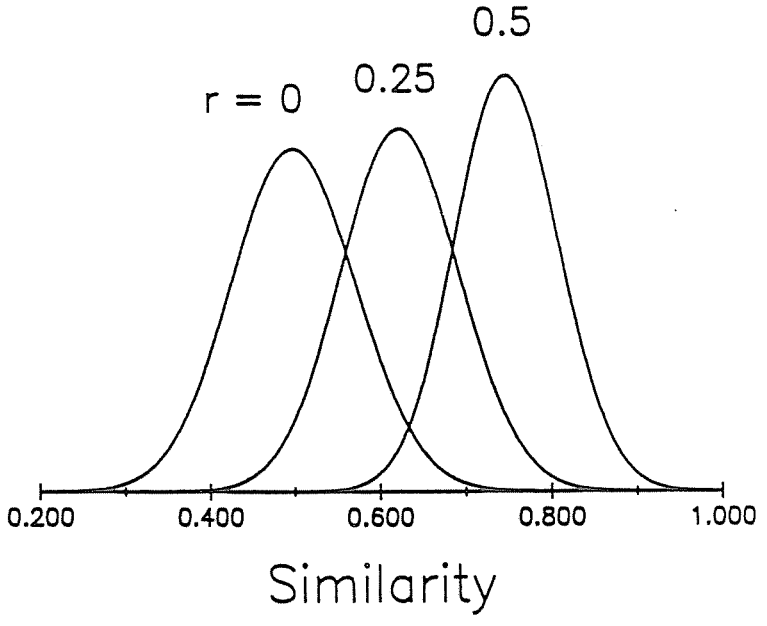

Figure 3. Expected distributions of similarity for individuals with $r = 0$ (nonrelatives), $r = 0.125$ (half-sibs or grandparent-grandchild), and $r = 0.5$ (full-sibs) obtained by use of Equations (15) and (22). It is assumed that the mean similarity of nonrelatives is $S_0 = 0.5$ and the mean number of bands is $\bar{n} = 40$.

Equation (15) to derive the expected distributions of S for different degrees of relatedness. Thus, knowing only $\bar{\theta}$ and $\bar{n}$ in advance, it is possible to get a good impression of the possible power of DNA-fingerprinting to resolve issues of relatedness (Fig. 3). For investigators involved in long-term research programs with specific populations, it may be useful to go a step further and directly establish base-line data on the empirical distributions of S for known types of relatives. Such distributions could be used in future investigations to test the hypothesis that a pair of unknown individuals has a similarity value consistent with a specific type of relationship.

Finally, it may be noted that a modification of Equation (24) can be used to generate an estimate of the degree of relatedness between mates:

$$r_m = \frac{\bar{S}_m - \bar{\theta}}{1 - \bar{\theta}}, \tag{24}$$

where $\bar{S}_m$ is the mean number of shared bands for mates. A simple test for consanguineous mating is to evaluate whether $\bar{S}_m$ and $\bar{\theta}$ are significantly different.

Discussion

The main point of this paper has been to show how information extracted from DNA-fingerprint analyses is related to conventional population genetic parameters. Most of the connections that have been pointed out make no assumptions about the gene-frequency distributions for the loci sampled. In principle, much more precise statements can be made if the distribution of genotypes is known since the exact distribution of band number and band-sharing can be expressed in terms of formulae of Chakraborty (1981). However, this possibility will not be realized as long as investigators rely on multilocus probes.

The data from multilocus DNA-fingerprinting studies are somewhat phenomenological in nature, and consequently they do not usually yield unbiased estimators of most of the measures of population structure that routinely appear in population genetic theory. Nevertheless, as noted above, they do provide reasonable first approximations in many cases. It is even possible that the magnitude of the bias from DNA-fingerprinting surveys is relatively small compared to the sampling error that results from more conventional surveys involving isozymes and single-locus RFLPs. That is a useful area for future theoretical and empirical investigation.

Acknowledgements
This work has been supported by NSF grant BSR 86-00487 and PHS grant R01 GM36827-01. Helpful comments were provided by R. Chakraborty, A. Jeffreys, and other conference participants.

References

Brookfield, J. F. Y. (1989) Analysis of DNA fingerprinting data in cases of disputed paternity. *IMA J. Math. Appl. Med. Biol.* 6: 111–131.

Burke, T., and Bruford, M. W. (1987) DNA fingerprinting in birds. *Nature* 327: 149–152.

Burke, T., Davies, N. B., Bruford, M. W., and Hatchwell, B. J. (1989) Parental care and mating behavior of polyandrous dunnocks *Prunella modularis* related to paternity by DNA fingerprinting. *Nature* 338: 249–251.

Chakraborty, R. (1981) The distribution of the number of heterozygous loci in an individual in natural populations. *Genetics* 98: 461–466.

Cohen, J. E. (1990) DNA fingerprinting for forensic identification: potential effects on data interpretation of subpopulation heterogeneity and band number variability. *Am. J. Hum. Genet.* 46: 358–368.

Georges, M., Lathrop, M., Hilbert, P., Marcotte, A., Schwers, Swillens, S., Vassart, G., and Hanset, R. (1990) On the use of DNA fingerprints for linkage studies in cattle. *Genomics* 6: 461–474.

Georges, M., Lequarré, A.-S., Castelli, M., Hanset, R., and Vassart, G. (1988) DNA fingerprinting in domestic animals using four different minisatellite probes. *Cytogenet. Cell. Genet.* 47: 127–131.

Gilbert, D. A., Lehman, N., O'Brien, S. J., and Wayne, R. K. (1990) Genetic fingerprinting reflects population differentiation in the California Channel Island fox. *Nature* 344: 764–767.

Gyllensten, U. B., Jakobsson, S., Temrin, H., and Wilson, A. C. (1989) Nucleotide sequence and genomic organization of bird minisatellites. *Nucleic Acids Res.* 17: 2203–2214.

Jeffreys, A. J., Brookfield, J. F. Y., and Semeonoff, R. (1985a) Positive identification of an immigration test-case using human DNA fingerprints. *Nature* 317: 818–819.

Jeffreys, A. J., and Morton, D. B. (1987) DNA fingerprints of dogs and cats. *Animal Genetics* 18: 1–15.

Jeffreys, A. J., Neumann, R., and Wilson, V. (1990) Repeat unit sequence variation in minisatellites: a novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. *Cell* 60: 473–485.

Jeffreys, A. J., Royle, N. J., Wilson, V., and Wong, Z. (1988) Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. *Nature* 332: 278–280.

Jeffreys, A. J., Wilson, V., and Thein, S. L. (1985b) Hypervariable 'minisatellite' regions in human DNA. *Nature* 314: 67–73.

Jeffreys, A. J., Wilson, V., and Thein, S. L. (1985c) Individual-specific 'fingerprints' of human DNA. *Nature* 316: 76–79.

Kuhnlein, U., Zadworny, D., Dawe, Y., Fairfull, R. W., and Gavora, J. S. (1990) Assessment of inbreeding by DNA fingerprinting: development of a calibration curve using defined strains of chickens. *Genetics* 125: 161–165.

Lander, E. S. (1989) DNA fingerprinting on trial. *Nature* 339: 501–505.

Lynch, M. (1988) Estimation of relatedness by DNA fingerprinting. *Mol. Biol. Evol.* 5: 584–599.

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

Lynch, M., and Crease, T. J. (1990) The analysis of population survey data on DNA sequence variation. *Mol. Biol. Evol.* 7: 377–394.

Malécot, G. (1948) Les Mathématiques de l'Hérédité. Masson et Cie. Paris.

Morton, E. S., Forman, L., and Braun, M. (1990) Extrapair fertilizations and the evolution of colonial breeding in purple martins. *Auk* 107: 275–283.

Nei, M. (1972) Genetic distance between populations. *Amer. Natur.* 106: 283–292.

Nei, M. (1987) Molecular evolutionary genetics. Univ. Columbia Press. New York.

Wetton, J. H., Carter, R. E., Parkin, D. T., and Walters, D. (1987) Demographic study of a wild house sparrow population by DNA fingerprinting. *Nature* 327: 147–152.

Wright, S. (1951) The genetical structure of populations. *Ann. Eugen.* 15: 323–354.