Analyzing molecular data for studies of genetic diversity

Fernando González-Candelas and Carmen Palacios
Departament de Genètica, Universitat de València, Valencia 46100, Spain

Introduction

One of the main goals in conservation is the preservation of genetic diversity. Traditionally, the study of genetic diversity has fallen within population genetics, which has focussed on measuring its extent in natural populations, in comparing levels of genetic diversity within and among populations and in making inferences on the nature and intensity of evolutionary processes from the observed patterns of genetic diversity. Hence, there is a long tradition as well as a wealth of conceptual tools in population genetics for analyzing, measuring and partitioning genetic diversity.

This review on methods for analyzing variation using molecular markers will start with a brief outline of the main population genetic concepts involved. These ideas were developed for simple situations, such as the one-locus two-alleles case, and were refined and generalized later. However, the main features are best understood by taking the simplest case which, in terms of a molecular marker, can be understood as an allozyme locus with only two alleles. In this situation, we are dealing with codominant markers, for which all possible genotypes (both homozygotes and the heterozygote) can be easily ascertained.

We will then move to the case of the richest possible markers in terms of the amount and quality of the information provided, DNA sequences. For these markers, it is possible to establish measurements of their evolutionary distance, which can further be used to refine the measurements of genetic diversity. Once the direct analysis of nucleotide sequences has been developed, we will consider other markers which provide indirect estimates of nucleotide divergence between alternative alleles, such as RFLPs and restriction site data. After that, we shall consider the complicating effects of using dominant markers, such as RAPDs and multilocus DNA-fingerprinting, for which the "presence" allele is dominant over the "absence" allele. Furthermore, use of these markers usually results in an unknown number of loci being analyzed simultaneously which introduces further complications. Finally, microsatellite markers will be considered, as these can be interpreted with or without reference to the evolutionary relatedness among alleles. Some computer programmes available for use in population genetics and analysis of molecular variation are listed in the Appendix to this paper.

The population genetics description of diversity

Genetic diversity can be measured, as can any other measurement of diversity, in different ways. One of the most commonly used ways defines diversity in a single locus as

$$D_l = 1 - \sum_i p_{li}^2 \qquad \text{Eq.1}$$

where p_{li} represents the frequency of the *i*-th allele at locus *l*. An average diversity for several (*L*) loci is given by

$$D = 1 - \frac{1}{L}\sum_{l} D_l = 1 - \frac{1}{L}\sum_{l}\sum_{i} p_{li}^2 \qquad \text{Eq.2}$$

This definition of diversity is closely related to the expected heterozygosity in a single locus for diploid organisms when populations are in Hardy-Weinberg equilibrium. Under these circumstances, the expected heterozygosity in a locus is given by:

$$H_l = 2\sum_{i \neq j} p_{li}p_{lj} = 1 - \sum_{l} p_{li}^2 = D_l \qquad \text{Eq.3}$$

This relationship provides an interesting solution to the problem of comparing levels of diversity between haploid organisms, where genetic diversity is readily defined but heterozygosity is not, and diploid organisms.

Following Weir (1990), it is possible to obtain a partition of the distribution of genetic diversity estimated directly from heterozygosity values in terms of an analysis of variance (ANOVA), taking into account the several levels at which heterozygosity can be defined. So the effects of subpopulations, of individuals within subpopulations, of the different loci and their interactions on the heterozygosity observed in a population can be easily obtained and tested for the significance of their relative contributions to the observed variability.

Deviations from Hardy-Weinberg equilibrium reflect on differences between observed and expected values of heterozygosity. These deviations, which can be due to many different causes, can be formulated in terms of inbreeding coefficients. So the genotypic frequencies in a two-allele locus, *P* and *Q* for homozygotes and *H* for heterozygotes, can be expressed (Wright 1931) as:

$$P = p^2 + fpq \qquad \text{Eq.4}$$

$$H = 2pq - 2fpq \qquad \text{Eq.5}$$

$$Q = q^2 + fpq \qquad \text{Eq.6}$$

where *f* is the inbreeding coefficient. Its sign and value reflect deviations from Hardy-Weinberg proportions. When *f* takes a positive value, there will be an excess of homozygotes and a lack of heterozygotes, as when endogamic reproduction occurs. Conversely, negative values of *f* are an indication of exogamy. An indirect estimate of the amount of inbreeding for a given locus is obtained from the observed proportion of heterozygotes, H_0, as

$$\hat{f} = 1 - \frac{H_0}{2\hat{p}\hat{q}} \qquad \text{Eq.7}$$

where $\hat{p}$ and $\hat{q}$ are the estimated gene frequencies of each allele. The corresponding sampling variance for a sample of size *n* is given by (Rasmussen 1964)

$$Var(\hat{f}) = \frac{1}{2n\hat{p}\hat{q}}\left\{(1-\hat{f})\left[2\hat{p}\hat{q}(1+\hat{f}) + \hat{f}(2-\hat{f})(1-2\hat{p})^2\right]\right\} \qquad \text{Eq.8}$$

It immediately follows that the amount of genetic diversity in a given locus is related to the level of inbreeding in that population.

The two major, non-selective causes that move natural populations of diploid organisms away from Hardy-Weinberg equilibrium are drift and inbreeding. Both factors act on reducing the amount of heterozygotes and increasing homozygosity, and this effect is further enhanced whenever populations become structured i.e. mating and dispersal take place only in the close neighbourhood of each individual. Wright (1951) introduced a method of describing genetic population structures of diploid organisms in terms of three *F*-statistics or allelic correlations.

In every subdivided population, there are at least three levels of complexity: individual organisms (*I*), subpopulations (*S*) and the whole population (*P*). These three levels are associated with three different measurements of heterozygosity:

H_I, which can be interpreted as the average heterozygosity of all the genes in a single individual or the probability of heterozygosity in any gene. H_I is the observed heterozygosity averaged over populations. If H_{il} represents the heterozygosity in a single locus in subpopulation *i* (Eq. 9) and *k* subpopulations and *L* loci are considered, then

$$H_{Il} = \frac{1}{k}\sum_{i=1}^{k} H_{il} \qquad H_I = \frac{1}{L}\sum_{l=1}^{L} H_{Il} \qquad \text{Eq.9}$$

H_S represents the heterozygosity level expected in a panmictic subpopulation. Hence, for a diallelic locus with gene frequencies p_i and q_i in subpopulation *i*, H_S always equals $2p_iq_i$. For *k* subpopulations, the average value of the H_S for each subpopulation is represented by $\bar{H}_S$.

H_T represents the expected heterozygosity of all subpopulations when pooled and mating is random in the pooled population. In this case, H_T is given by $2p_0q_0$, being p_0 the average gene frequency across subpopulations.

Wright developed these ideas for the diallelic case, grouping all but the most frequent allele into a single class. An explicit multiallelic form was presented by Nei (1987), although the notation employed was slightly different. Nei uses G_{ST} to refer to themultiallelic form of F_{ST} developed by Wright which, accordingly should be reserved only for the diallelic case (Nei 1987). So, if p_{iX} is the frequency of the *i*-th allele in subpopulation *X*, then:

$$H_S = 1 - \sum_{i=1}^{k} p_{iX}^2 , \qquad \text{Eq.10}$$

and if $\bar{p}_i$ is the average frequency of the *i*-th allele over subpopulations, then

$$H_T = 1 - \sum_{i=1}^{k} \bar{p}_i^2 \qquad \text{Eq.11}$$

The inbreeding coefficient measures the reduction in individual heterozygosity due to deviations from random mating in the local populations. This inbreeding coefficient is represented by F_{IS} and is given by:

$$F_{IS} = \frac{\bar{H}_S - H_I}{H_S} \qquad \text{Eq.12}$$

The effects of population subdivision can be quantified by means of the fixation index F_{ST}, which is the reduction in the heterozygosity in a subpopulation due to nonrandom mating with respect to the total population. F_{ST} is given by:

$$F_{ST} = \frac{H_T - \overline{H}_S}{H_T} \qquad \text{Eq.13}$$

An alternative interpretation of F_{ST} in its diallelic version is as the ratio between the expected and observed variances of gene frequency considered among all subpopulations. So:

$$F_{ST} = \frac{\sigma_p^2}{p_0 q_0} \qquad \text{Eq.14}$$

The following relationship holds for *F*-statistics:

$$(1 - F_{IS})(1 - F_{ST}) = (1 - F_{IT}) \qquad \text{Eq.15}$$

The estimation of *F*-statistics by mere substitution in the previous equations of the relevant parameters by their observed values does not necessarily lead to better estimates, especially with small sample sizes. Ideally, the estimates should be corrected for the effects of sampling a limited number of individuals in a limited number of subpopulations. Several corrections have been proposed (Wright 1968; Curie-Cohen 1982; Nei and Chesser 1986; Weir and Cockerham 1984; Nei 1986) although further difficulties arise with their application.

Although several statistical tests have been proposed for testing the significance of each of these *F*-statistics (Li and Horvitz 1953; Brown 1970; Workman 1970), Cockerham (1969, 1973) pioneered the development of a system for analyzing *F*-statistics in an adequate context for hypothesis testing. Cockerham worked on the relationship between *F*-statistics and components of variance under an ANOVA framework. This work was further developed by Weir and Cockerham (1984) and Long (1986).

Excoffier *et al.* (1992) have introduced an alternative method for partitioning genetic variance at different levels which is based on an extension of the works of Cockerham (1973), Long (1986) and Long *et al.* (1987) on the allelic correlation among demes. This method uses the usual setup of the analysis of variance on a transformed measure of distance between different genetic variants (usually haplotypes, see below) obtaining an evolutionary metric distance. The method is implemented in the program WINAMOVA (Excoffier *et al.*).

Summarizing, population geneticists usually evaluate genetic diversity by means of observed and estimated amounts of heterozygosity, and they compare and partition this variation among different hierarchical population levels by means of Wright's *F*-statistics and related quantities.

There are, however, alternate ways of analyzing the same basic information. Another way to study the level of genetic differentiation among populations is by asking how similar they are. It is generally considered that genetic distance increases with time of divergence from a common population. But this requires a genetic model that specifies those genetic processes, such as migration and drift, that make populations diverge. Many genetic distance measures have been proposed (Reynolds 1981; Nei 1987) since Cavalli-Sforza and Edward's (1967) attempt to relate their distance measure to the evolutionary changes of gene frequencies among populations. Some of the distance measures used in genetic studies are geometric distances that do not consider special features of evolutionary processes. This is the case of Mahalanobis' (1936), Bhattacharyya's (1946) and Rogers' (1972) distances, to name a few. More appropriate measures when dealing with genetic data have been developed by Nei (1972, 1973). Two of them are especially relevant for this review, the minimum genetic distance (D_m) and the standard genetic distance (D).

Let p_{iX} and p_{jY} represent the frequency of allele i in a given locus in population X and the frequency of allele j in the same locus in population Y. Suppose we take random allele from each population and compare them. The probability that both alleles are identical is given by:

$$j_{XY} = \sum_i p_{iX} p_{iY} \qquad \text{Eq.16}$$

and they will be different with probability $1\text{-}j_{XY}$. When the alleles are different, there is at least one codon or nucleotide, depending on what marker is being used, difference between them. Therefore $d'_{XY}=1\text{-}j_{XY}$ gives the minimum number of differences between both populations. However, when there is a polymorphism, two alleles randomly sampled from one population will not always be identical. Hence, we need to correct for these intrapopulational differences. For each population, the intrapopulation measure of differentiation is given by:

$$d_X = 1 - \sum_i p_{iX}^2 \qquad \text{Eq.17}$$

which equals the expected heterozygosity (and the gene diversity) for that locus in that population. Therefore the net minimum number of differences between two populations is given by:

$$d_{XY} = d'_{XY} - \frac{d_X + d_Y}{2} = \sum_i \frac{(p_{iX} - p_{iY})^2}{2} \qquad \text{Eq.18}$$

In practice, d varies from one locus to another, and in order to estimate the difference between two populations, the average of d over all loci must be taken. This average is known as *minimum genetic distance* and is given by:

$$D_m = D_{XY(m)} - \frac{D_{X(m)} + D_{Y(m)}}{2} \qquad \text{Eq.19}$$

where $D_{XY(m)} = 1 - J_{XY}$, $D_{X(m)} = 1 - J_X$, and $D_{Y(m)} = 1 - J_Y$, and J_{XY}, J_X and J_Y are the averages of j_{XY}, j_X and j_Y over all loci. The main disadvantage of using the minimum genetic distance is that it can seriously underestimate the difference between pairs of populations. However, it can be used for the study of the maintenance of polymorphism among populations (Chakraborty 1974).

When changes occur independently at every position in the genome, the mean number of net substitutions is given by:

$$D = -\log I, \qquad \text{Eq.20}$$

where

$$I = \frac{J_{XY}}{\sqrt{J_X J_Y}} \qquad \text{Eq.21}$$

This is known as the *standard genetic distance*. *I* takes value 1 when the two populations have identical gene frequencies in all loci and 0 when they share no alleles. Because of this property, *I* itself has been used as a measure of the genetic similarity between populations, and it is known as the *genetic identity*. Nei (1987) discusses the estimation procedure and sampling properties of several estimators of genetic distances.

Generally, more than two populations of any species are being analyzed, and all possible pairwise genetic distances (or identities) have been estimated. In these cases, the information provided in the corresponding distance matrix can be used to simultaneously analyze the relationships among all the populations. This is usually accomplished by means of different multivariate techniques (Manly 1986), of which clustering methods are most popular. The two most commonly used clustering methods are UPGMA (Sokal and Michener 1958; Sneath 1973) and neighbour-joining (Saitou and Nei 1987). UPGMA is an ultrametric method and should provide accurate clusters when distances are linearly proportional to the amount of divergence, for instance under constancy of evolutionary rates, whereas neighbour-joining is a very robust method which is gaining widespread use because of its relative independence of assumptions (Swofford and Olsen 1990; Nei 1991).

Using nucleotide sequence data

There are two different quantities for measuring the amount of genetic variation at the DNA level: the average number of pairwise nucleotide differences and the number of segregating (polymorphic) sites among a sample of sequences.

The number of segregating sites, *S*, is the number of sites which are occupied by at least two different nucleotides. The average number of pairwise nucleotide differences among DNA sequences is defined as:

$$d = \frac{2\sum_{i<j} d_{i,j}}{n(n-1)} \qquad \text{Eq.22}$$

where $d_{i,j}$ is the number of nucleotide differences between sequences *i* and *j*, and *n* is the number of DNA sequences sampled from a population. When the number of DNA sequences studied is large, it is advisable to use heterozygosity instead of *d*. At a site *i*, heterozygosity is defined as:

$$H_i = 1 - \sum_{j=1}^{4} p_j^2 \qquad \text{Eq.23}$$

where p_j is the relative frequency of nucleotide j (j = 1, 2, 3, 4 corresponding to nucleotides A, T, C and G) and an unbiased estimate is given by (Tajima 1993):

$$\hat{H}_i = \frac{n}{n-1}\left(1-\sum_{j=1}^{4} p_{ij}^2\right) \tag{Eq.24}$$

where p_{ij} is the observed frequency of nucleotide j at site i and n is the number of nucleotide sequences. The following relationship can be shown between heterozygosity and the average number of nucleotide differences among the n sequences studied

$$d = \sum_{i=1}^{m} H_i \tag{Eq.25}$$

where m is the number of nucleotide sites in the DNA sequence.

Both S and d depend on the length of the nucleotide sequence (m) and the amount of DNA polymorphism per site can be used instead, simply by dividing any of those measurements by m. The new measures correspond to the average number of nucleotide differences per site (S/m) and to the average heterozygosity per site (H/m). Tajima (1963) provides an excellent review of both measures under different evolutionary scenarios.

In order to compare the amount of variation at several levels using sequence data, it is necessary to define the average number of nucleotide differences per site between two sequences, an amount also known as *nucleotide diversity*. It can be defined as:

$$\pi = \sum_{i,j} x_i x_j \pi_{ij} \tag{Eq.26}$$

It can immediately be shown that π can be estimated by:

$$\hat{\pi} = \frac{n}{n-1}\sum_{i<j} \hat{x}_i \hat{x}_j \pi_{ij} = \frac{2\sum_{i<j}\pi_{ij}}{n(n-1)} = \frac{S}{m} \tag{Eq.27}$$

An alternative estimate of π was proposed by Nei and Miller (1990):

$$\hat{\pi}_p = \frac{\sum_{i=1}^{m} h_i}{m} \tag{Eq.28}$$

which is equivalent to the average heterozygosity per site.

The average proportion of nucleotide differences between n_X sequences from population X and n_Y sequences from population Y can be calculated as:

$$\bar{p}_{XY} = \frac{\sum_{i=1}^{m} h_{XY_i}}{m} \tag{Eq.29}$$

where h_{XYi}, the proportion of nucleotide differences at site i, is given by:

$$h_{XY_i} = 1 - \sum_{j=1}^{4} p_{Xij} p_{Yij} \qquad \text{Eq.30}$$

An approximate value of d'_{XY} can de obtained as:

$$\hat{d}'_{XY} = -\frac{3}{4} \ln\left(1 - \frac{4}{3} \bar{p}_{XY}\right) \qquad \text{Eq.31}$$

and the interpopulation component of the nucleotide differentiation between populations X and Y is then given by:

$$\hat{d}_{XY} = \hat{d}'_{XY} - \frac{\hat{d}_X + \hat{d}_Y}{2} \qquad \text{Eq.32}$$

whose expected value for a pair of populations that diverged t years ago is $2\lambda t$, being λ the rate of nucleotide substitution per site per year.

When several populations are analyzed, Lynch and Crease (1990) devised a method for partitioning nucleotide diversity into intra- and interpopulation components which is analogous to F_{ST} at the DNA level. Their method is based on the average number of nucleotide substitutions per site between pairs of sequences sampled both from each population and from all possible pairs of populations. The intrapopulation component is estimated as:

$$\hat{v}_X = \frac{2}{n_X(n_X - 1)} \sum_{i,j} n_{iX} n_{jX} \hat{d}_{ij} \qquad \text{Eq.33}$$

where n_X is the total number of individuals sampled in population X, n_{iX} and n_{jX} are the number of those individuals with haplotypes i and j, respectively, and d_{ij} is the estimated nucleotide distance between those haplotypes. The combined estimate of intrapopulation differentiation is given by:

$$\hat{v}_w = \frac{\sum_{i=1}^{n_p} \hat{v}_i}{n_p} \qquad \text{Eq.34}$$

being n_p the number of populations studied. The combined estimate of interpopulation differentiation is obtained as:

$$\hat{v}_b = \frac{2 \sum_{X<Y} \hat{v}_{XY}}{n_p(n_p - 1)} \qquad \text{Eq.35}$$

The analogue to the indices of population structure (Wright's F_{ST}) proposed by Lynch and Crease (1990) at the nucleotide level is:

$$N_{ST} = \frac{\hat{v}_b}{\hat{v}_b + \hat{v}_w} \qquad \text{Eq.36}$$

which is the ratio between the average genetic distance between genes from different populations and the average global genetic distance. The extreme values of N_{ST}, 0 and 1, are indication of null and complete population subdivision, respectively.

The approximate sampling variance of N_{ST} is given by:

$$Var(\hat{N}_{ST}) = \left(\frac{\hat{N}_{ST}}{\hat{v}_w + \hat{v}_b}\right)^2 \left[\left(\frac{\hat{v}_w}{\hat{v}_b}\right)^2 Var(\hat{v}_b) - 2\left(\frac{\hat{v}_w}{\hat{v}_b}\right) Cov(\hat{v}_w, \hat{v}_b) + Var(\hat{v}_w)\right] \qquad \text{Eq.37}$$

If we assume, as a first approximation, that N_{ST} is normally distributed then the statistic:

$$D = \frac{N_{ST}^2}{Var(N_{ST})} \qquad \text{Eq.38}$$

will be chi-square distributed with 1 degree of freedom under the null hypothesis of no population subdivision.

The analysis of restriction data

In order to analyze genetic diversity using restriction data, it is necessary to introduce a few previous ideas. In the first place, it is important to distinguish between restriction fragment data, for which only the size of the generated fragments is available, and restriction site data, for which the precise location of a recognition sequence for a restriction enzyme is known. These types of data are not adequate for comparing sequences which have diverged considerably, but both are usually acceptable for studying intraspecific variation. In order to compare levels of genetic diversity within and among populations from restriction data, it is necessary to previously estimate the number of nucleotide substitutions between any pair of sequences. Several methods, both for restriction fragment and restriction site data are reviewed in Nei (1987).

For restriction site data, let S denote the probability that two sequences, X and Y, share the same recognition sequence at a given site. This value can be described by:

$$S = (1-p)^r \qquad \text{Eq.39}$$

where p is the probability that the sequences do not share a nucleotide in a given position and r represents the length of the recognition sequence. The probability p is related to the expected number of substitutions per site, d, according to:

$$p = \frac{3}{4}\left[1 - e^{\left(-\frac{4}{3}d\right)}\right] \qquad \text{Eq.40}$$

If the rate of nucleotide substitution per site and per year is λ, then d is also given by $d = 2\lambda t$. Consequently, it is possible to estimate d if the value of S is known. Whenever nucleotide divergence is relatively small ($d < 0.25$), as is certainly the case for sequences from the same and very closely related species, S is usually approximated by (Nei and Li 1979; Kaplan and Risko 1981; Li 1981):

$$S = e^{-2r\lambda t} \qquad \text{Eq.41}$$

The maximum likelihood estimator of S (Nei and Tajima 1983) is given by:

$$\hat{S} = \frac{2m_{XY}}{m_X + m_Y} \qquad \text{Eq.42}$$

with variance given by:

$$Var(\hat{S}) = \frac{\hat{S}(1-\hat{S})(2-\hat{S})}{m_X + m_Y} \qquad \text{Eq.43}$$

where m_X and m_Y are the number of restriction sites in sequence X and Y, respectively, and m_{XY} is the number of restriction sites shared by both sequences.

Once an estimate of S is available, it is possible to estimate the proportion of nucleotide differences, p, by:

$$\hat{p} = 1 - \sqrt[r]{\hat{S}} \qquad \text{Eq.44}$$

and the estimate of d follows immediately:

$$\hat{d} = -\frac{3}{4}\ln\left(1 - \frac{4}{3}\hat{p}\right) \qquad \text{Eq.45}$$

When $d < 0.25$, it is possible to use the above approximation (Nei and Li 1979) which leads to the estimate:

$$\hat{d}_1 = -\frac{\ln \hat{S}}{r} \qquad \text{Eq.46}$$

In the preceding derivation, it has been assumed that a single enzyme has been used. If several enzymes with the same length of the corresponding recognition sequences are used, it is possible to use the above expression simply by taking summations over all the enzymes. However, if several enzymes with different lengths in their recognition sequences are employed, then it is convenient to follow the method proposed by Nei and Miller (1990) in order to weigh the data obtained with each enzyme class. When values of d have been estimated for each class of restriction enzyme according to Eqs. 45 or 46 above, then a combined estimate of d for all the enzymes is given by:

$$\hat{d} = \frac{\sum_k \hat{m}_k r_k \hat{d}_k}{\sum_k \hat{m}_k r_k} \qquad \text{Eq.47}$$

where $\hat{m}_k$ is the average number of bands for the k-th class of enzymes, r_k is the length of the sequence recognized by the class of enzymes and $\hat{d}_k$ is the corresponding estimated nucleotide divergence.

This estimate of nucleotide divergence can be extended to be an estimate of interpopulation nucleotide divergence for all possible pairs of populations X and Y simply by considering the nucleotide divergence estimated for all possible pairs of sequences one from each subpopulation. This leads to:

$$\hat{S}_{XY} = \frac{2\sum_{i,j} m_{X_iYj}}{\sum_{i,j} m_{X_i(Y_j)} + \sum_{i,j} m_{(X_i)Y_j}} \qquad \text{Eq.48}$$

where m_{XiYj} represents the number of restriction sites shared by the i-th sequence from population X and the j-th sequence from population Y, whereas ; $m_{Xi(Yj)}$ and $m_{(Xi)Yj}$ are the number of restriction sites in sequences i and j form subpopulations X and Y, respectively. This estimate can be obtained for each different class of restriction enzymes, and these estimates can be combined, similarly as above, into the estimate:

$$\hat{d}'_{XY} = \frac{\sum_k \hat{m}_k r_k \hat{d}_{XY_k}}{\sum_k \hat{m}_k r_k} \qquad \text{Eq.49}$$

where:

$$\hat{m}_k = \frac{\hat{m}_{X_k} + \hat{m}_{Y_k}}{2} \qquad \text{Eq.50}$$

This estimate of interpopulation divergence includes both an interpopulation component, due to differences in the frequency of the different sequences in both subpopulations, as well as an intrapopulation component, due to variation within each subpopulation. If we are interested merely in interpopulation divergence, then the interpopulation component of the above estimate of nucleotide divergence between both subpopulations can be estimated according to:

$$\hat{d}_{XY} = \hat{d}'_{XY} - \frac{\hat{d}_X + \hat{d}_Y}{2} \qquad \text{Eq.51}$$

Unfortunately, the method developed by Nei and Li (1979) to estimate the sampling variance of d_{XY} cannot be applied in this situation (Nei and Miller 1990) and resampling estimates, by jackknifing or bootstrapping, must be obtained.

When only restriction fragment length data are available, the estimate of nucleotide divergence, as the average number of nucleotide substitutions per site, between a pair of sequences can be obtained as:

$$\hat{d} = -\frac{2}{r}\ln \hat{G} \qquad \text{Eq.52}$$

where $G = e^{-r\lambda t}$ is the probability that no nucleotide substitution has occurred at a restriction site and is related to the proportion of shared fragments (F) by means of:

$$F = \frac{G^4}{3-2G} \qquad \text{Eq.53}$$

and F can be estimated from:

$$\hat{F} = \frac{2m_{XY}}{m_X + m_Y} \qquad \text{Eq.54}$$

Nei (1987) has proposed an iteration procedure to estimate G, once an estimate for F has been obtained from Eq. 54.

Similar expressions for the estimates of intra- and interpopulation nucleotide divergence can be obtained. González-Candelas *et al.* (1995) have recently derived an approximate expression for the sampling variance of d in this situation.

There are two main ways of measuring the amount of polymorphism in a given population. The haplotype diversity, h, was defined by Nei and Tajima (1981) as:

$$H = 1 - \sum_{i=1} p_i^2 \qquad \text{Eq.55}$$

where p_i is the frequency of the i-th haplotype. An estimate of H can be obtained as:

$$\hat{H} = \frac{2n}{2n-1}\left(1 - \sum_i \hat{p}_i^2\right) \qquad \text{Eq.56}$$

This definition is equivalent to that of heterozygosity or genic diversity described above.

A second way to measure the amount of polymorphism is by the average number of differences in the restriction sites between pairs of randomly chosen haplotypes. This number is given by:

$$v = \sum_{i,j} p_i p_j v_{ij} \qquad \text{Eq.57}$$

and an unbiased estimate is obtained as:

$$\hat{v} = \frac{n}{n-1} \sum_{i,j} \hat{p}_i \hat{p}_j v_{ij} \qquad \text{Eq.58}$$

However, these two measures of polymorphism depend on the length of the DNA fragment studied by restriction analysis. This can be avoided by using a measure of variation at the nucleotide level. As most polymorphisms in restriction sites are due to nucleotide substitutions, this can be used to estimate nucleotide diversity, defined as:

$$d=\sum_{i,j} p_i p_j d_{ij} \qquad \text{Eq.59}$$

where d_{ij} is the proportion of nucleotide differences between haplotypes i and j and p_i and p_j are their respective population frequencies. We have already seen how d can be estimated from restriction site or restriction fragment data. An unbiased estimate of nucleotide diversity is then given by:

$$\hat{d}=\frac{n}{n-1}\sum_{i,j} \hat{p}_i \hat{p}_j d_{ij} \qquad \text{Eq.60}$$

Once an estimate of d for any given population is available, it is possible to extend the procedure in order to analyze the amount of divergence among populations. Let d_X represent the average number of nucleotide substitutions between a pair of haplotypes randomly chosen in population X. This number can be estimated as:

$$\hat{d}_X=\frac{n_X}{n_X-1}\sum_{i,j} \hat{p}_i \hat{p}_j d_{ij} \qquad \text{Eq.61}$$

where n_X represents the sample size in population X. The average number of substitutions between pairs of haplotypes randomly chosen one from each population X and Y can be estimated from:

$$\hat{d}'_{XY}=\sum_{i,j} \hat{p}_i \hat{p}_j d_{ij} \qquad \text{Eq.62}$$

when the i-th and j-th haplotypes have been sampled from populations X and Y, respectively. As before, in this measure both an intra- and an interpopulation component of variability are included. The net amount of nucleotide substitutions between these two populations is then estimated by:

$$\hat{d}_{XY}=\hat{d}'_{XY}-\frac{\hat{d}_X+\hat{d}_Y}{2} \qquad \text{Eq.63}$$

If the rate of nucleotide substitution per site and per year between two populations that diverged t years ago is given by λ, then the expected value of d_{XY} is:

$$d_{XY}=2\lambda t \qquad \text{Eq.64}$$

Hence, in order to compute the time since divergence between two populations (t), it is necessary to subtract from d_{XY} the average nucleotide difference between polymorphic alleles at the instant of separation. The sampling variance of d_A is given by:

$$Var\left(\hat{d}_{XY}\right)=Var\left(\hat{d}'_{XY}\right)+\frac{1}{4}\left[Var\left(\hat{d}_X\right)+Var\left(\hat{d}_Y\right)\right]-\left[Cov\left(\hat{d}'_{XY}.\hat{d}_X\right)+Cov\left(\hat{d}'_{XY}.\hat{d}_y\right)\right] \qquad \text{Eq.65}$$

Nei (1987) gives expressions for the corresponding variances and covariances in the above expression.

In the above derivations, it has been assumed that changes in recognition sites occur only once, and hence no provision has been made for superposition of substitutions. This is a good approximation as long as divergence between sequences is rather small, but when the number of changes increases, the approximation no longer holds. Then it becomes necessary to correct for the probability of several substitutions on the same recognition site. The corresponding expression was derived by Nei and Tajima (1983):

$$\delta_{XY} = -\frac{3}{4}\ln\left(1-\frac{4}{3}d'_{XY}\right) \qquad \text{Eq.66}$$

When several populations are analyzed, the method proposed by Lynch and Crease (1990) for partitioning nucleotide diversity the into intra- and interpopulation components described in the preceding section can be readily applied to both restriction site or restriction fragment data.

The analysis of RAPD data

The analysis of RAPD data has been hampered by the failure of usual methods to correct for the inability of detecting genotypes with dominant markers. This can result in a serious underestimation of the actual level of genetic diversity (Clark and Lanigan 1993). Recently, two methods for overcoming this difficulty and thus enabling the use of RAPD data have been proposed (Clark and Lanigan 1993; Lynch and Milligan 1994).

The method proposed by Clark and Lanigan (1993) uses the frequency of the absence of a fragment in a population sample as an estimate of the population frequency of recessive heterozygotes (q^2) and then uses this value to correct for the relative detectability of individuals who have one versus two copies of a fragment. Once this correction has been taken into account, data are treated in a very similar way to that already described for restriction fragment data. This is feasible if the following assumptions are met:

1. Amplification of a fragment is dependent on the primer hybridizing to the flanking sequences. In other words, if one single substitution is present in the sequence complementary to the primer this will not hybridize and the corresponding fragment will not be amplified.
2. Individuals being analyzed should diverge in less than 0.10, because the model does not allow for multiple hits. This assumption is usually met in population studies.
3. Sizes of the bands can be determined accurately and all different bands can be told apart from each other.
4. Different bands represent independent loci in linkage equilibrium. Incidentally, this is one of the less realistic assumptions in this model.
5. Populations are panmictic and samples are taken randomly.
6. Hardy-Weinberg equilibrium can be assumed for genotypic proportions.

As above, let *P* be the probability that no mutation has occurred at a primer site since the divergence from the common ancestor of two sequences. If *F* is the expected proportion of fragments that remain unchanged then, as for restriction fragments, Nei and Li (1979) showed the approximate relationship:

$$F = \frac{P^4}{3-2P} \qquad \text{Eq.67}$$

From the number of bands shared by two individuals (m_{XY}), and those present in individuals X (m_X) and Y (m_Y), the following estimate of F can be obtained:

$$\hat{F} = \frac{2m_{XY}}{m_X + m_Y} \qquad \text{Eq.68}$$

The expected nucleotide divergence between two sequences is $d = 2\lambda t$ if λ is the rate of nucleotide substitution per site and per year, and since $P = \exp(-r\lambda t)$, it is possible to estimate d from the relation:

$$\hat{d} = -\frac{2}{r}\ln\hat{P} \qquad \text{Eq.69}$$

The preceding estimate of d is the nucleotide divergence for a pair of haploid individuals. If several individuals from each of two populations are examined, it is possible to estimate the interpopulational nucleotide divergence following Nei and Miller (1990):

$$\hat{F}_{XY} = \frac{2\sum_{i,j} m_{X_iY_j}}{n_Y\sum_i m_{X_i} + n_X\sum_j m_{Y_j}} \qquad \text{Eq.70}$$

where m_{XiYj} is the number of bands shared by individual i from population X and individual j from population Y, m_{Xi} is the number of bands scored in individual i from population X and n_X and n_Y are the number of individuals sampled in the corresponding populations. Sample sizes are used to weigh the number of bands scored in each population in order to make the number of within and between-population comparisons the same.

The correction for dominance is based on the assumption of Hardy-Weinberg equilibrium. The expected heterozygosity under Hardy-Weinberg equilibrium is $2pq$. The conditional heterozygosity of band i in an individual from population X, given that band is observed, is defined as:

$$H_{X(i)} = \frac{2pq}{p^2 + 2pq} \qquad \text{Eq.71}$$

The numbers m_X, m_Y and m_{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:

$$m_X = \sum_i\left[4\left(1-H_{X(i)}\right)^2 + 4H_{X(i)}\left(1-H_{X(i)}\right) + H^2_{X(i)}\right] \qquad \text{Eq.72}$$

$$m_Y = \sum_i\left[4\left(1-H_{Y(i)}\right)^2 + 4H_{Y(i)}\left(1-H_{Y(i)}\right) + H^2_{Y(i)}\right]$$

$$m_{XY} = \sum_i\left[4\left(1-H_{X(i)}\right)\left(1-H_{Y(i)}\right) + 2\left(1-H_{X(i)}\right)H_{Y(i)} + 2H_{X(i)}\left(1-H_{Y(i)}\right) + H_{X(i)}H_{Y(i)}\right]$$

After the weighted values of m_{XY}, m_X and m_Y are tallied for all bands and pairs of individuals, F and d can be calculated from Equations 67 - 70.

Nei and Takezaki (1994) proposed a modified estimate of F values based on the frequencies of each band instead of their direct count, and on taking a geometric rather than an arithmetic average for comparing the shared bands with the bans present in the common ancestor. This leads to:

$$\hat{F} = \frac{\sum_i p_{X_i} p_{Y_i}}{\sqrt{\sum_i p_{X_i}^2 \sum_i p_{Y_i}^2}} \qquad \text{Eq.73}$$

where p_{Xi} represents the frequency of the i-th DNA fragment in population X. This frequency, for diploid organisms and in populations in Hardy-Weinberg equilibrium, can be estimated from:

$$p_{X_i} = 1 - \sqrt{Q_{X_i}} \qquad \text{Eq.74}$$

where Q_{Xi} is the frequency of individuals lacking the i-th fragment in population X.

From the estimate of nucleotide divergence, d, obtained using Eq. 69 it is possible to evaluate the interpopulational component of the total diversity as:

$$d_{XY} = d'_{XY} - \frac{d_X + d_Y}{2} \qquad \text{Eq.75}$$

When several different primers are used and sample sizes differ from one primer to another and/or primers have different lengths, it is possible to combine the different data into one single estimate by using an approach similar to Nei and Miller's (1990):

$$\hat{d}_{XY} = \frac{\sum_k \bar{n} r_k \hat{d}_{XY_k}}{\sum_k \bar{n} r_k} \qquad \text{Eq.76}$$

where $\bar{n}$ is the average number of individuals assayed in each population X and Y, r_k is the length of the k-th primer and $\hat{d}_{XYk}$ is the corresponding estimate of interpopulation divergence for that primer.

Lynch and Milligan (1994) have adopted a different approach for analyzing population structure using RAPDs. They simply assume that alleles from different loci do not comigrate to the same position in the gel, that the researcher is capable of matching bands from different lanes within and among gels, and that each locus can be treated as a two-allele system, with a presence and an absence allele. They adopt the following estimate for the gene frequency, q, of the null allele at one locus:

$$\hat{q} = \frac{\sqrt{\hat{x}}}{1 - \frac{Var(\hat{x})}{8\hat{x}^2}} \qquad \text{Eq.77}$$

where x is the frequency of null homozygotes. This is an asymptotically unbiased estimator of q and has lower bias than the correction proposed by Clark and Lanigan (1993).

Once gene frequencies have been estimated, it is possible to estimate gene diversity within a population. The usual measure of gene diversity:

$$H_{X_i} = 2p_{Xi}q_{Xi} = 1 - \sum_i p_{Xi}^2 \tag{Eq.78}$$

which is the probability that two genes randomly chosen from population X differ at the i-th locus, is equivalent to the expected herterozygosity under Hardy-Weinberg equilibrium. An estimator of this quantity is given by

$$\hat{H}_{X_i} = 2\hat{q}_{Xi}\hat{p}_{Xi} + 2Var\left(\hat{q}_{Xi}\right) \tag{Eq.79}$$

whose sample variance is approximately:

$$Var\left(\hat{H}_{X_i}\right) = 4\left(1 - 2\hat{q}_{Xi}\right)^2 Var\left(\hat{q}_{Xi}\right) \tag{Eq.80}$$

If L loci have been sampled in population X, the average gene diversity in this population is:

$$\hat{H}_X = \frac{1}{L}\sum_{i=1}^{L}\hat{H}_{X_i} \tag{Eq.81}$$

and if n populations have been sampled, the average within-population gene diversity can be estimated by:

$$\hat{H}_W = \frac{1}{n}\sum_{X=1}^{n}\hat{H}_X \tag{Eq.82}$$

Expressions for the corresponding sample variances can be found in Lynch and Milligan (1994).

The heterozygosity between populations X and Y at the i-th locus can be estimated by:

$$\hat{H}'_{XY_i} = \hat{q}_{Xi} + \hat{q}_{Yi} - 2\hat{q}_{Xi}\hat{q}_{Yi} \tag{Eq.83}$$

If there is no population subdivision, the gene frequencies in all populations are the same, so $\hat{H}'_{XY_i} = \hat{H}_{Xi} = \hat{H}_{Yi}$, and the interpopulational component of diversity can be estimated as usual by:

$$\hat{H}_{XY_i} = \hat{H}'_{XY_i} - \frac{\hat{H}_{X_i} + \hat{H}_{Y_i}}{2} \tag{Eq.84}$$

Averaging over all loci, the estimated mean gene diversity between populations X and *Y* is:

$$\hat{H}_{XY} = \frac{1}{L}\sum_{i=1}^{L} \hat{H}_{XY_i}, \qquad \text{Eq.85}$$

and the mean between population gene diversity can be obtained by averaging over all possible pairs of populations:

$$\hat{H}_B = \frac{2\sum_{X<Y} \hat{H}_{XY}}{n(n-1)} \qquad \text{Eq.86}$$

Lynch and Milligan (1994) propose an asymptotically unbiased estimate of F_{ST} by using:

$$\hat{F}_{ST} = \frac{\hat{H}_B}{\hat{H}_B + \hat{H}_W} \frac{1}{\left[1+\frac{\hat{H}_B Var(\hat{H}_W) - \hat{H}_W Var(\hat{H}_B) + (\hat{H}_B - \hat{H}_W) Cov(\hat{H}_B, \hat{H}_W)}{\hat{H}_B(\hat{H}_B + \hat{H}_W)^2}\right]} \qquad \text{Eq.87}$$

The expression for the variances and covariances in the above equation can be found in Lynch and Milligan (1994). There are two packages of freeware programmes for the analysis of RAPD data, RAPDISTANCE (Armstrong *et al.* 1995) and RAPDIS (Dopazo 1995).

An alternative approach for estimating F-statistics from RAPDs data has been used by Huff *et al.* (1993) and Peakall *et al.* (1995). They have used the already described analysis of molecular variance (Excoffier *et al.* 1992) implemented in the programme WINAMOVA in order to estimate population differentiation statistics. This procedure is especially useful when more than two population levels have to be considered.

One of the potential uses of RAPDs is their capability for providing estimates of the relatedness among individuals in the population. This measure is based on the expectation that related individuals will have more similar genotypes than nonrelatives, and hence the fraction of loci for which two individuals are identical should increase with the degree of relatedness. An estimate for the relatedness, *r*, between individuals *a* and *b* using data at the *i*-th locus is given by (Lynch and Milligan 1994):

$$\hat{r}_{ab_i} = \frac{S_{ab_i} - \hat{\theta}_i}{1-\hat{\theta}_i} + \frac{1-S_{ab_i}}{[1-\hat{\theta}_i]^3} Var(\hat{\theta}_i) \qquad \text{Eq.88}$$

where $S_{abi} = 1$ or 0 denotes whether individuals *a* and *b* have the band at the *i*-th locus or not, θ_i is the probability that two nonrelatives have matching phenotypes at the locus, and:

$$\hat{\theta}_i = 1 - 2Q_i(1-Q_i)\left(1-\frac{1}{N}\right) \qquad \text{Eq.89}$$

$$Var(\hat{\theta}_i) = \frac{4Q_i(1-Q_i)(2Q_i-1)^2}{N} \qquad \text{Eq.90}$$

A more accurate estimate is obtained by averaging over all *L* loci:

$$\hat{r}_{ab} = \frac{1}{L}\sum_{i=1}^{L}\hat{r}_{ab_l} \qquad \text{Eq.91}$$

Such an analysis is only applied to polymorphic loci. Nevertheless, due to the overlap between the distributions of similarity for RAPDs for different degrees of relatedness, the utility of relatedness estimation using RAPDs is rather limited (Lynch and Milligan 1994).

Analyzing DNA fingerprinting data

Hypervariable minisatellite DNA, when analyzed by Southern hybridization following restriction digestion, produce DNA fingerprints. These fingerprints usually correspond to several loci which share a core sequence, hence showing all at once in a single gel. These loci usually exhibit large allelic diversity, and hence only on a few occasions will individuals from exogamous populations sampled at random show exactly the same DNA fingerprint pattern.

In order to develop similarity estimates, potential indicators of the relative level of population homozygosity, a few assumptions on the technical ability of the researcher have to be made (Lynch 1990; Lynch and Crease 1990). First, it is assumed that the DNAs of individuals to be compared are run in close lanes and/or with adequate controls so that errors on the assignment of identity to pairs of fragments are minimized. Second, it is assumed that all individuals are sampled at random from the population. Third, it is assumed that all comigration of non-allelic markers can be resolved either by differences in band intensity or by some other means. Fourth, marker loci are assumed unlinked and in Hardy-Weinberg equilibrium within and among loci. And last, the same set of homologous loci is tested in all individuals.

Similarity is usually defined as the fraction of shared bands. For two individuals, *x* and *y*, it can be defined as the number of common fragments (n_{xy}) divided by an estimate of the number of fragments in any individual:

$$S_{xy} = \frac{2n_{xy}}{n_x + n_y} \qquad \text{Eq.92}$$

It is necessary to relate this index with a population genetic parameter such as the identity by state between pairs of individuals and population homozygosity. The identity in state for two individuals can be defined as 100% for pairs of individuals AA-AA or Aa-Aa and as 50% for pairs such as AA-Aa or Aa-Aa′. The expected genotypic identity in state for a panmictic population is:

$$E(I) = \frac{\sum_{k}\sum_{i} P_{ki}^2 + P_{ki}^2\,(1\text{-}P_{ki})^2}{L.} \qquad \text{Eq.93}$$

where p_{ki} is the frequency of the i-th allele at the k-th locus and L is the number of loci. Alternatively, the identity in state can be defined from the standpoint of gametes taken at random from two individuals. Under panmictic mating, the expected gametic identity in state is equivalent to population homozygosity:

$$E(H)=\frac{\sum_k\sum_i p_{ki}^2}{L} \qquad \text{Eq.94}$$

Jeffreys *et al.* (1985) and Lynch (1988) showed that:

$$E(S)=\frac{\sum_k\sum_i p_{ki}^2(2-p_{ki})}{L} \qquad \text{Eq.95}$$

Hence, the similarity index is always a biased estimator by excess both of I and H. Lynch (1988) developed the sampling theory for the similarity index. When large numbers of polymorphic loci are sampled, the sampling variance of the average population similarity can be directly estimated from the observed data as:

$$Var(\bar{S})=\frac{NVar(S_{xy})+2N'Cov(S_{xy},S_{xz})}{N^2} \qquad \text{Eq.96}$$

where N is the total number of similarity measures used and N' is the number of pairs of those measures shared by an individual. When average similarities from different populations are being compared and there is no certainty that the same loci have been sampled in all populations or we are interested in making inferences on properties of the whole genome from the sampled loci, it is convenient to use the following expression that takes into account both the effect of sampling on the studied loci and the error due to sampling a finite number of loci:

$$Var'(S_{xy})=\frac{2\bar{S}(1-\bar{S})(2-\bar{S})}{\bar{n}(4-\bar{S})} \qquad \text{Eq.97}$$

where $\bar{n}$ is the average number of bands present in any individual.

A measure of interpopulation similarity corrected for intrapopulation similarity is given by:

$$\bar{S}_{ij}=1+\bar{S}'_{ij}-\frac{\bar{S}_i+\bar{S}_j}{2} \qquad \text{Eq.98}$$

where $\bar{S}_i$ is the average similarity for individuals belonging to the i-th populations and $\bar{S}'_{ij}$ is the average similarity between pairs of individuals randomly sampled from populations i and j. As the similarity index is not an unbiased estimator of population homozygosity, caution should be taken when using it for estimating the usual measure of population subdivision, Wright's F-statistics. Nevertheless, if the biases corresponding to $\overline{S}_{ij}, \overline{S}_i$ and $\overline{S}_j$ are approximately equal, then they cancel out in the above expression for $\overline{S}_{ij}$. Consequently, $\overline{D}_{ij} = 1 - \overline{S}_{ij}$ is an unbiased estimator of the interpopulation genetic diversity.

Let D_b denote the average of $\overline{D}_{ij}$ for all i, j and let D_w denote the average value of $1 - S_i$. Then:

$$F' = \frac{D_b}{D_b + D_w} \qquad \text{Eq.99}$$

provides a downwards biased estimate of population subdivision.

An unbiased estimate of the average heterozygosity for a system with L loci is given by (Stephens *et al.* 1992):

$$\hat{H} = \frac{\frac{2n}{2n-1}\sum_{i=1}^{L}\left(1-\sum_{j=1}^{A_i} p_{ij}^2\right)}{L} = \frac{2n}{2n-1}\left(1-\frac{\sum_{k=1}^{A} p_k^2}{L}\right) \qquad \text{Eq.100}$$

where A_i is the number of alleles in the i-th locus and p_{ij} is the estimated frequency of the j-th allele in the i-th locus. The second equality represents the lack of importance for the estimation of heterozygosity of the specific distribution of alleles throughout loci.

As every band or allele in a fingerprint is effectively dominant, it is necessary to estimate the allele frequency (p_k) from the frequency with which the k-th band appears (S_k). Assuming Hardy-Weinberg equilibrium for the genotypes, then:

$$p_k = 1 - \sqrt{1 - s_k} \qquad \text{Eq.101}$$

The individual p_k estimates can be summed up to provide an estimate of L (Gilbert *et al.* 1990). An improved estimate of heterozygosity is obtained when monomorphic and polymorphic bands are considered separately. Let L_M represent the number of monomorphic loci and A_P that of polymorphic bands, such that $A_P = A - L_M$, being A the total amount of bands. Heterozygosity will be at a maximum when all allele frequencies in polymorphic loci are uniform, i.e. when for each allele frequency $p_{ij} = 1/A_{ij} = L_P / A_P$. Then the estimate of the maximum heterozygosity will be:

$$H_{max} = \frac{2n}{2n-1}\,\frac{L_P\left(1-\frac{L_P}{A_P}\right)}{L_M + L_P} \qquad \text{Eq.102}$$

and the value of L_P is given by

$$L_P = \sqrt{L_M A} - L_M. \qquad \text{Eq.103}$$

Microsatellites and SSR loci

The main difficulty posed by microsatellite loci for their use in the evaluation of genetic distance is their relatively high mutation rate. This makes it difficult to adopt any of the two main mutation models used in population genetics, the infinite alleles or the stepwise mutation model. There is still uncertainty as to whether allele sizes are unconstrained or whether there are certain limits to the number of repeats present (Estoup *et al.* 1995; Garza *et al.* 1995; Meyer *et al.* 1995). Assuming a stepwise mutation model, Slatkin (1995) and Goldstein *et al.* (1995) have recently proposed a distance measure for microsatellite alleles. The distance between two alleles is a simple transformation of the number of repeat units. The within population measure of distance is obtained as the average sum of squares of the differences in number of repeats between alleles:

$$S_{Wj} = \frac{2}{2n(2n-1)} \sum_{i<i'} \left(a_{ij} - a_{i'j}\right)^2 \qquad \text{Eq.104}$$

where a_{ij} is the allele size of the *i*-th copy ($i = 1,\ldots,2n$) in the *j*-th population ($j = 1,\ldots,d_s$). The average within population distance S_w from Slatkin is equivalent to D_0 from Goldstein *et al.* (1995):

$$S_W = \frac{1}{d_s} \sum_{j=1}^{d_s} S_{Wj} \qquad \text{Eq.105}$$

In order to estimate the average distance between all possible pairs of alleles, it is necessary to define the between population component, S_b as:

$$S_B = \frac{2}{(2n)^2 d_s\left(d_s - 1\right)} \sum_{j<j'} \sum_{i<i'} \left(a_{ij} - a_{i'j'}\right)^2 \qquad \text{Eq.106}$$

which is equivalent to D_1 of Goldstein *et al.* (1995). The global distance is obtained by a weighted average of the intra- and interpopulation components:

$$\bar{S} = \frac{2n-1}{2nd_s - 1} S_W + \frac{2n\left(d_s - 1\right)}{2nd_s - 1} S_B \qquad \text{Eq.107}$$

where the coefficients represent the probability of choosing two different copies of the locus from the same and from different populations, respectively. In practice, it is easier to compute S_w and $\bar{S}$. directly from the variances of allele sizes, as S_w is twice the average of the variances of allele size within each population and $\bar{S}$ is twice the estimated variance of allele size in the collection of populations together. MICROSAT is a programme that can be used for computing these distances.

Given that S_w and $\bar{S}$ are proportional to the within-population and total variances, the fraction:

$$R_{ST} = \frac{\bar{S} - S_W}{\bar{S}} \qquad \text{Eq.108}$$

has the same properties for microsatellite loci that follow the stepwise mutation model as F_{ST} has for allozyme loci. R_{ST} is simply the fraction of the total variance in allele size that is due to interpopulation differences."

An extension of this method, incorporating the analysis of microsatellite data into an ANOVA framework, has been recently proposed by Michalakis and Excoffier (1995). In this method, the partition of genetic variance at different levels is achieved by means of an analysis of molecular variance, as described above, by using the programme WINAMOVA.

An alternative distance measure, the shared allele distances D_{AS} (Chakraborty and Jin 1993) has been advocated by Estoup *et al.* (1995) for use with microsatellite data. This distance is computed by averaging the values over all loci at pairs of individuals. For each locus, the distance is 1 if both individuals have the same genotype, 0 if they have no allele in common and 0.5 if they share only one allele. With the use of this distance, it is possible to group individuals by any of the different methods of clustering (see above). The programme MICROSAT can also be used to compute D_{AS} distances.

Shriver *et al.* (1995) have proposed the use of a stepwise weighted genetic distance measure (D_{SW}), which is an extension of Nei's minimum genetic distance. This measure has several advantages over minimum and standard genetic distances when applied to loci evolving via a stepwise mutation mechanism. Let p_{Xi} represent the allele frequency of the *i*-th allele in population *X*. The proposed distance weighs the probability that two alleles are different when randomly sampled from one or two populations by the absolute value of the difference in steps (number of repeats for tandem repeat loci) between the two alleles. That is:

$$d_{XW} = \sum_{i \neq j} p_{Xi} p_{Xj} \delta_{ij}$$ Eq.109

$$d_{YW} = \sum_{i \neq j} p_{Yi} p_{Yj} \delta_{ij}$$ Eq.110

$$d_{XYW} = \sum_{i \neq j} p_{Xi} p_{Yj} \delta_{ij}$$ Eq.111

where:

$$\delta_{ij} = |i - j|$$ Eq.112

D_{SW} can then be defined as:

$$D_{SW} = d_{XYW} - \frac{d_{XW} + d_{YW}}{2}$$ Eq.113

As in the case of Nei's distance measures, D_{SW} can be estimated by means of the unbiased estimates of d_{XW}, d_{YW} and d_{XYW}, given by:

$$\hat{d}_{XW} = \frac{n_X}{n_X - 1} \sum_{i \neq j} \hat{p}_{Xi} \hat{p}_{Xj} \delta_{ij}$$ Eq.114

$$\hat{d}_{YW} = \frac{n_Y}{n_Y - 1} \sum_{i \neq j} \hat{p}_{Yi} \hat{p}_{Yj} \delta_{ij}$$ Eq.115

$$\hat{d}_{XYW} = \sum_{i \neq j} \hat{p}_{Xi} \hat{p}_{Yj} \delta_{ij}$$ Eq.116

where n_X and n_Y are the number of chromosomes sampled from populations X and Y, respectively. For the estimation of D_{SW} from multilocus data, the averages over all loci of the above estimators can be used.

Acknowledgements

This work has been supported by grant PB93-0350 from DGICYT (Ministerio de Educación y Ciencia, España).

References

Armstrong, J., A. Gibbs, R. Peakall and G. Weiller. 1995. RAPDistance programs; Version 1.03 for the analysis of patterns of RAPD fragments.

Bhattacharyya, A. 1946. On a measure of divergence between two multinomial populations. Sankhya 7:401-406.

Brown, A.H.D. 1970. The estimation of Wright's fixation index from genotypic frequencies. Genetica 41:399-406.

Cavalli-Sforza, L.L. and A.W.F. Edwards. 1967. Phylogenetic analysis: Models and estimation procedures. Am. J. Hum. Genet. 19:233-257.

Chakraborty, R. 1974. A note on Nei's measure of gene diversity in a substructured population. Hummangenetik 21:85-88.

Chakraborty, R. and L. Jin. 1993. A unified approach to study hypervariable polymorphisms: statistical considerations of determining relatedness and population distances. Pp. 153-175 *in* DNA Fingerprinting: State of the Science (S.D.J. Peña, R. Chakraborty, T.J. Epplen and A.J. Jeffreys, eds.). Birkhauser Verlag, Basel.

Clark, A.G. and C.M.S. Lanigan. 1993. Prospects for estimating nucleotide divergence with RAPDs. Molecular Biology and Evolution 10:1096-1111.

Cockerham, C.C. 1969. Variance of gene frequencies. Evolution 23:72-84.

Cockerham, C.C. 1973. Analysis of gene frequencies. Genetics 74:679-700.

Curie-Cohen, M. 1982. Estimates of inbreeding in a natural population: A comparison of sampling properties. Genetics 100:339-358.

Dopazo, J. 1995. RAPDIS (In preparation).

Estoup, A., L. Garnery, M. Solignac and J.M. Cornuet. 1995. Microsatellite variation in honey bee (*Apis mellifera* L.) populations: hierarchical genetic structure and test of the infinite allele and stepwise mutation models. Genetics 140:679-695.

Excoffier, L., P.E. Smouse and J.M. Quattro. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479-491.

Gilbert, D.A., Y.A. Reid, M.H. Gail, D. Pee, C. White, R.J. Hay and S.J. O'Brien. 1990. Application of DNA fingerprints for cell-line individualization. Am. J. Hum. Genet. 47:499-514.

Goldstein, D.B., A. Ruiz Linares, L.L. Cavalli-Sforza and M.W. Feldman. 1995. An evaluation of genetic distances for use with microsatellite loci. Genetics 139:463-471.

González-Candelas, F., S.F. Elena and A. Moya. 1995. Approximate variance of nucleotide divergence between two sequences estimated from restriction fragment data. Genetics 140:1443-1446.

Huff, D.R., R. Peakall and P.E. Smouse. 1993. RAPD variation within and among natural populations of outcrossing buffalograss [*Buchloë dactyloides* (Nutt.) Engelm.]. Theor. Appl. Genet. 86:927-934.

Jeffreys, A.J., V. Wilson and S.L. Thein. 1985. Hypervariable "minisatellite" regions in human DNA. Nature 314:67-73.

Jin, L. and J.W.H. Ferguson. 1990. Neighbor-joining tree and UPGMA tree software.

Kaplan, N. and K. Risko. 1981. An improved method for estimating sequence divergence of DNA using restriction endonuclease mappings. J. Molec. Evolution 17:156-162.

Li, C.C. and D.G. Horvitz. 1953. Some methods of estimating the inbreeding coefficient. Am. J. Hum. Genet. 95:107-117.

Li, W.-H. 1981. A simulation study if Nei and Li's model for estimating DNA divergence from restriction enzyme maps. J. Molec. Evolution 17:251-255.

Long, J.C. 1986. The allelic correlation structure of Gainj- and Kalam-speaking people. I. The estimation and interpretation of Wright's F-statistics. Genetics 112:629-647.

Long, J.C., P.E. Smouse and J.W. Wood. 1987. The allelic correlation structure of Gainj. and Kalam-speaking people. II. The genetic distance between population subdivisions. Genetics 117:273-283.

Lynch, M. 1988. Estimation of relatedness by DNA fingerprinting. Molec. Biol. and Evolution 5:584-599.

Lynch, M. 1990. The similarity index and DNA fingerprinting. Molec. Biol. and Evolution 7:478-484.

Lynch, M. and B.G. Milligan. 1994. Analysis of population genetic structure with RAPD markers. Molec. Ecol. 3:91-99.

Lynch, M. and T.J. Crease. 1990. The analysis of population survey data on DNA séquence variation. Molec. Biol. and Evolution 7:377-394.

Mahalanobis, P.C. 1936. On the generalized distance in statistics. Proc. Natl. Inst. Sci. India 2:49-55.

Manly, J.B.F. 1986. Multivariate Statistical Methods. A Primer. Chapman and Hall, London.

Michalakis, Y. and L. Excoffier. 1995. A generic estimation of population subdivision using distances between alleles with specieal interest to microsatellite loci. Genetics (In press).

Miller, J.C. 1990. A program for computing distances between phylogenetic groups based on restriction-site or fragment data.

Nei, M. 1972. Genetic distance between populations. The American Naturalist 106:283-292.

Nei, M. 1973. Analysis of gene diversity in subdivided populations. Proc. Nat. Acad. Sci., USA 70:3321-3323.

Nei, M. 1973. The theory and estimation of genetic distance. Pp. 45-54 *in* Genetic Structure of Populations (N.E. Morton, ed.). University Press of Hawaii, Honolulu.

Nei, M. 1986. Definition and estimation of fixation indices. Evolution 40:643-645.

Nei, M. 1987. Molecular Evolutionary Genetics. Columbia University Press, New York.

Nei, M. and F. Tajima. 1981. DNA polymorphism detectable by restriction endonucleases. Genetics 97:145-163.

Nei, M. and F. Tajima. 1983. Maximum likelihood estimation of the number of nucleotide substitutions from restriction sites data. Genetics 105:207-217.

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 N. Takezaki. 1994. Estimation of genetic distances and phylogenetic trees from DNA analysis. Proc. 5th World Cong. Genet. Appl. Livestock Prod. 21:405-412.

Nei, M. and R.K. Chesser. 1983. Estimation of fixation indices and gene diversities. Am. J. Hum. Genet. 47:253-259.

Nei, M. and W.H. Li. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Nat. Acad. Sci. USA 76:5269-5273.

Peakall, R., P.E. Smouse and D.R. Huff. 1995. Evolutionary implications of allozyme and RAPD variation in diploid populations of dioecious buffalograss *Buchloë dactyloides*. Molec. Ecol. 4:135-147.

Rasmussen, D.I. 1964. Blood group polymorphism and inbreeding in natural populations of the deer mouse *Peromyscus maniculatus*. Evolution 18:219-229.

Raymond, M. and F. Rousset. 1995. GENEPOP (V. 1.2): A population genetics software for exact tests and ecumenicism. J. Heredity (in press).

Reynolds, J. 1981. Genetic Distance and Coancestry. North Carolina State University, Raleigh, NC, USA.

Rogers, J.S. 1972. Measures in genetic similarity and genetic distance. Studies in Genetics VII. University of Texas Publ. 7213:145-153.

Rohlf, F.J. and D.E. Slice. 1992. NTSYS-pc.

Rozas, J. and R. Rozas. 1995. DnaSP, DNA sequence polymorphism: an interactive program for estimating Population Genetics parameters from DNA sequence data. Computer Application in Biosciences (in press).

Saitou, N. and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. J. Molec. Evolution 4:406-425.

Shriver, M.D., L. Jin, E. Boerwinkle, R. Deka, R.E. Ferrell and R. Chakraborty. 1995. A novel measure of genetic distance for highly polymorphic tandem repeat loci. Molec. Biol. and Evolution 12:914-920.

Slatkin, M. 1995. A measure of population subdivision based on microsatellite allele frequencies. Genetics 139:457-462.

Sneath, P.H.A. and R.R. Sokal. 1973. Numerical Taxonomy. W.H. Freeman, San Francisco.

Sokal, R.R. and C.D. Michener. 1958. A statistical method for evaluating systematic relationships. Univ. Kansas Sci. Bull. 28:1409-1438.

Stephens, J.C., D.A. Gilbert, N. Yuhki and S.J. O'Brien. 1992. Estimation of heterozygosity for single-probe multilocus DNA fingerprints. Molec. Biol. and Evolution 9:729-743.

Tajima, F. 1993. Measurement of DNA polymorphism. Pp. 37-59 *in* Mechanisms of Molecular Evolution (N. Takahata and A.G. Clark, eds.). Sinauer, Sunderland.

Weir, B.S. 1990. Genetic Data Analysis. Sinauer Associates Inc. Sunderland.

Weir, B.S. and C.C. Cockerham. 1984. Estimating F statistics for the analysis of population structure. Evolution 38:1358-1370.

Workman, P.L. and J.D. Niswander. 1970. Population studies on southwestern Indian tribes. II. Local genetic differentiation in the Papago. Am. J. Hum. Genet. 22:24-49.

Wright, S. 1931. Evolution in Mendelian populations. Genetics 16:97-159.

Wright, S. 1951. The genetical structure of populations. Ann. Eugen. 15:323-354.

Wright, S. 1978. Evolution and the Genetics of Populations. IV. Variability within and among Natural Populations. University of Chicago Press, Chicago.

Appendix

Computer programmes for use in population genetics and analysis of molecular variation

BIOSYS-1 (Swofford and Selander 1981, 1989)
Programme for the analysis of allelic variation in population genetics.

DnaSP (Rozas and Rozas 1995)
This is interactive programme for estimating population genetics parameters from DNA sequence data.
http://www.ebi.ac.uk
ftp://ftp.ebi.ac.uk

GENEPOP (Raymond and Rousset 1995)
GENEPOP is a population genetic software package, able to perform two major tasks:

1) It computes exact tests: for Hardy-Weinberg equilibrium, for population differentiation and for genotypic disequilibrium among pairs of loci.
2) It converts the input GENEPOP file to formats used by other programmes, like Biosys (Swofford and Selander 1981), Diploid (Weir 1990), Linkdos (Garnier-Gere and Dillman 1992) and M. Slatkin's (1993) isolation-by-distance programme (the last three programmes are also provided with GENEPOP, with the authorization of their authors).

ftp:// ftp.cefe.cnrs-mop.fr/ pub/msdos/genepop

MICROSAT (Goldstein *et al.* 1995)
Programme for computing distance measures with microsatellite data.
http://lotka.stanford.edu/research/microsat.html
ftp://lotka.stanford.edu/pub/Programs/microsat.c

NEIGHBOR (Jin and Ferguson 1990)
NJTREE, UPGMA and TDRAW is a group of software used for creating neighbour-joining trees or UPGMA trees.

NTSYS (Rolfe and Slyce 1992)
General package for multivariate analysis in population and evolutionary biology.

RAPDIS (Dopazo 1995)
Programme for the analysis of RAPD data.
http://www.tdi.es/

RAPDISTANCE (Armstrong *et al.* 1995)
RAPDistance Programmes; Version 1.03 for the Analysis of Patterns of RAPD Fragments.
ftp://life.anu.edu.au/pub/RAPDistance
http://life.anu.edu.au/molecular/software/rapd.html

RESTSITE (Miller 1990)
Programme for computing distances between phylogenetic groups based on restriction-site or fragment data.

WINAMOVA (Excoffier *et al.* 1992)
Programme for the analysis of molecular variance.
ftp://acasun1.unige.ch/pub/comp/win/amova
http://acasun1.unige.ch/LGB/Software/Windoze/amova