Heredity (2002) 88, 62–65

www.nature.com/hdy

# How many alleles per locus should be used to estimate genetic distances?

ST Kalinowski
*Conservation Biology Division, Northwest Fisheries Science Center, National Marine Fisheries Service, N.O.A.A., Seattle, WA 98112, USA*

As more microsatellite loci become available for use in genetic surveys of population structure, population geneticists are able to select loci to use in population structure surveys. This study used computer simulations to investigate how the number of alleles at loci affects the precision of estimates of four common genetic distances. This showed that equivalent results could be achieved by examining either a few loci with many alleles or many loci with a few alleles. More specifically, the total number of independent alleles appears to be a good indicator of how precise estimates of genetic distance will be.
*Heredity* (2002) **88,** 62–65. DOI: 10.1038/sj/hdy/6800009

**Keywords:** genetic distance; estimation; alleles; sampling; population genetics

## Introduction

Analysis of selectively neutral molecular markers has become a common method for inferring the evolutionary history of populations and species. Technological advances have enabled population and conservation geneticists to describe increasingly complex and subtle genetic relationships. However, molecular data remains expensive and many population level studies are limited by the amount of data that can be collected. Efficient study design remains important in order to maximize the ability of genetic data to describe genetic relationships between populations. Increasingly often, researchers have the luxury of selecting a set of loci to use in a study from a larger number of loci that have been previously characterized. This is particularly true for microsatellite loci. One of the most common criteria used to select loci is the number of alleles present. I shall call this the allele number of a locus. Loci with more alleles are generally thought to produce more precise estimates of genetic distances than loci with few alleles, especially for closely related populations. However, loci with a large number of alleles can be difficult to score and take up more space on electrophoretic gels. This space issue is usually not a problem when only one locus is run per gel, but it becomes a critical consideration when multiple loci are run on each gel. Unfortunately, there have been few guidelines for balancing these two contradictory concerns. In this investigation, I examine the relationship between allele number and the coefficient of variation of four popular genetic distances: the $D_A$ distance (Nei *et al*, 1983), the chord distance, $D_C$ (Cavalli-Sforza and Edwards, 1967), the standard genetic distance of Nei, $D_S$ (Nei, 1972, 1978), and the Weir and Cockerham estimator of $F_{ST}$, $\theta$ (Weir and Cockerham, 1984).

Each of these genetic distances has unique evolutionary and statistical properties (see Nei, 1987 for a review). Given a few simple assumptions, such as random mating and constant population size, the genetic distance between two halves of a population instantly split into two completely isolated new populations will initially increase linearly with time. For example, $D_S$ between two such population fragments will be equal to

$$D_S \approx 2\mu t$$

where $t$ is the number of generations the populations have been isolated, and $\mu$ is the mutation rate of the loci examined. In this case, $F_{ST}$ increases approximately linearly with time, provided $t$ is small

$$F_{ST} \approx \frac{t}{2N_e}$$

where $N_e$ is the effective size of each population. The expectations of $D_A$ and $D_C$ have not been expressed as functions of isolation time or other evolutionary variables, however they will initially increase linearly with time. All of these genetic distances are equal to zero when populations have the same allele frequencies. The maximum value of $D_A$, $D_C$, and $F_{ST}$ is equal to 1.0. $D_S$ will have a value of infinity when two populations do not share any alleles. The rate at which $D_A$, $D_C$, and $D_S$ increase with time is proportional to mutation rate. Therefore, each of these three genetic distances is expected to be higher at loci with many alleles than loci with fewer alleles.

*Correspondence: S Kalinowski, Conservation Biology Division, Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, 2725 Montlake Boulevard East, Seattle, WA 98112, USA. E-mail: Steven.Kalinowski@noaa.gov*
Received 3 April 2001; accepted 6 August 2001

## Methods

An analytical evaluation of the coefficient of variation for most genetic distances would be formidably complex, so I have relied on computer simulation to examine a simple model of population bifurcation (a similar simulation program available for general use is described by Excoffier *et al*, 2000). In this model, a randomly mating population having an effective population size of $N_e$ diploid individuals is instantaneously split into two completely isolated populations, each also having $N_e$ individuals. The populations are assumed to remain completely isolated until samples are collected $t$ generations after fragmentation. Gene trees for each locus were simulated using standard coalescent techniques (see Hudson, 1990 for review). The mutation rate for each locus was obtained by selecting a number, $u$, from the interval [–7, –2] and using $10^u$ as the mutation rate for that locus. Two mutational models were used: infinite alleles and single step mutation.

The infinite alleles model of mutation assumes that each mutation is unique. The single step model of mutation assumes that each allele can be represented as a sum, and that each mutation either adds or subtracts one from that sum. Mutation increasing the number of repeat units was assumed to be as likely as mutation decreasing the number of repeat units. All mutations were assumed to change the number of repeat units by a single step and no bounds were placed on the number of repeat units possible at simulated loci. Large numbers of loci were simulated and loci with 2, 3, 4, 8, 16, and 33 alleles were selected for analysis. Loci not having one of these numbers of alleles were dropped from analysis. All samples consisted of 100 diploid individuals from each population. The number of loci in the samples was varied from 2 to 32. Three effective population sizes ($N_e = 500$, 5000, and 50000) and three divergence times ($t = 50$, 500 and 5000) were examined. All combinations of these four parameters was examined (number of loci, number of alleles, $N_e$, $t$). Four commonly used genetic distances were estimated from the data: the $D_A$ distance (Nei *et al*, 1983), the chord distance, $D_C$ (Cavalli-Sforza and Edwards, 1967), the standard genetic distance of Nei, $D_S$ (Nei, 1978), and the Weir and Cockerham estimator of $F_{ST}$, $\theta$ (Weir and Cockerham, 1984).

The coefficients of variation for each of these genetic distances were estimated from the data by dividing the standard deviation of the estimates by the average estimate. Contour plots showing the coefficient of variation for data sets with different numbers of loci and varying numbers of alleles per locus were created with SigmaPlot 2000.

## Results and Discussion

Simulated data showed that highly polymorphic loci provided better estimates of genetic distances than less polymorphic loci (Figure 1). This trend was evident for the entire range of parameters examined ($N_e = 500$, 5000, 5000, $t = 50$, 500, 5000, number of alleles = 2, 4, 8, 16, 33). When the amount of population divergence was small, $t/N_e \leqslant 0.1$, the total number of independent alleles examined appeared to be a good indicator of the coefficient of variation of estimates of genetic distances (the number of independent alleles at a locus is equal to the total number of alleles at the locus minus one, and the total number of independent alleles is the sum of the number of independent alleles at each locus). For example, 16 loci each having two independent alleles had approximately the same coefficient of variation for each of the genetic distances as two loci having 16 independent alleles each. The ratio $t/N_e$ appeared to determine the relationship between number of alleles, number of loci, and the coefficient of variation. For example, the coefficient of variation of genetic distances for populations of 500 individuals separated for 50 generations was identical to the coefficient of variation of estimates of genetic distances between populations of 5000 individuals separated for 500 generations.

These results are in good agreement with an analysis of the Sanghvi (1953) genetic distance made by Foulley and Hill (1999). The Sanghvi distance is not used often for describing population structure, but it has tractable mathematical properties and has been shown to estimate phylogenies effectively (Takezaki and Nei, 1996). Foulley and Hill showed analytically that the coefficient of variation of the Sanghvi distance is approximately proportional to the sum of the number of independent alleles at each locus in the sample.

When divergence time was substantial, ie $t/N_e$ was 1.0 or greater, the relationship between the coefficient of variation of genetic distances and the number of independent alleles observed at small to moderate divergence times broke down, especially for highly polymorphic loci. This is not especially problematic, for the utility of these genetic distances to quantify genetic differences between highly differentiated populations is limited. Keep in mind that both genetic drift and mutation lead to differentiation. Both $D_A$ and $D_C$ approach their maximum value of 1.0 when mutation rate and divergence time are high. This results in these statistics having a low coefficient of variation when divergence time and polymorphism are high (Figure 1), but decreases their ability to describe the length of population separation. $D_S$ loses its utility when polymorphism is high, divergence time is long, and few loci are scored. In this case, samples from each population often share no alleles and $D_S$ is undefined. For example, about half of the loci having 33 alleles had no alleles in common in populations of 5000 individuals after 5000 generations. Lastly, $\theta$ has two undesirable properties when divergence time and polymorphism is high. It asymptotically approaches a maximum value, and this value is inversely proportional to the amount of polymorphism present in the populations (eg, Hedrick, 1999).

Of the four distances examined, $D_A$ and $D_C$ exhibited the strongest equivalence of alleles within and between loci (Figure 1). $D_S$ and $\theta$ did not fit this trend as closely. For both of these distance measures, adding more loci decreased the coefficient of variation faster than increasing the number of alleles per locus. For example, eight loci with two independent alleles produced better estimates of $D_S$ than two loci with eight independent alleles.

Mutation mechanism did not appear to have a strong affect upon the coefficient of variation when divergence time was short. When the length of population isolation was short, $t/N_e \leqslant 0.01$, the coefficient of variation of estimates of genetic distances appeared to be a function of the total number of independent alleles examined for both types of mutation mechanisms. When the length of

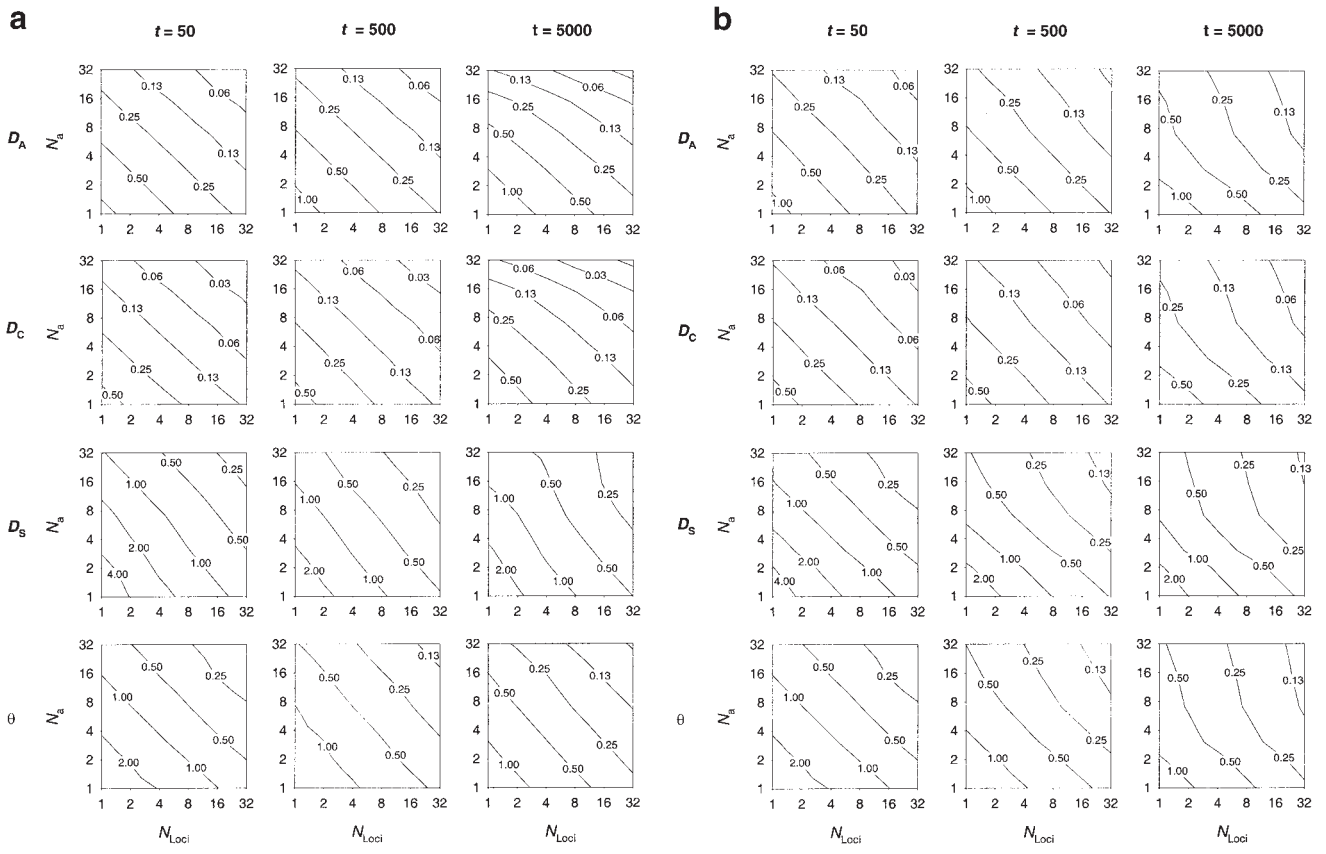


**Figure 1** The coefficient of variation of estimates of four genetic distances ($D_A$, $D_C$, $D_S$ and θ) for samples containing different numbers of loci and different numbers of alleles per locus. The two populations compared had a constant effective size of 5000 and had been isolated for $t = 50$, 500, or 5000 generations. Results for loci with infinite alleles mutation (**a**) and stepwise mutation (**b**) are shown.

isolation was longer, $t/N_e = 0.1$, this relationship was approximately true, but broke down to some extent at loci with stepwise mutation and high mutation rates. When the length of population isolate was quite long, $t/N_e \geqslant 1.0$, the contrast between mutation mechanisms was greatest (Figure 1). However, for both mutation mechanisms, the equivalence of alleles observed at low divergence times broke down at loci with high mutation rates (ie, those with greater than eight alleles) but not at loci with lower mutation rates.

Increasing allele number was associated with decreased coefficients of variation for each of the four genetic distances studied. The standard error of these statistics, however, behaved differently. The standard error of $D_S$, $D_A$, and $D_C$ increased as allele number increased (Figure 2). The coefficient of variation of these statistics decreased only because the average value of these genetic distances increased faster than the standard error (Figure 2). The expected value of θ is relatively insensitive to how much variation is present at loci and its standard error decreased as allele number increased.

The equivalent utility of alleles within and across loci for estimating genetic distances described here is significant because it demonstrates that study design for estimating genetic distances is flexible as long as the amount of divergence is not great. There is no requirement to examine either highly polymorphic loci or large numbers of loci. The only requirement is that a sufficient number of alleles be examined.

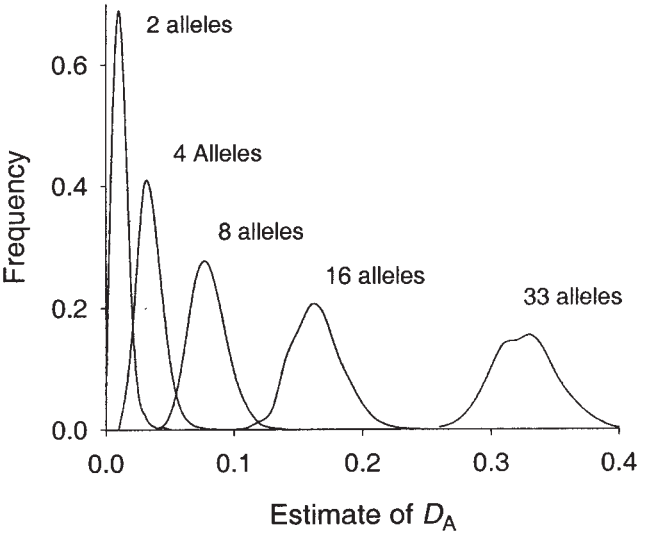


**Figure 2** Distributions of estimates of $D_A$ between two populations of 5000 individuals separated for 500 generations. Each estimate is based on eight simulated loci. A log-uniform distribution of mutation rates was used to simulate mutation rates and each mutation was assumed to create a new allele. Each line is a spline curve of the proportions of estimates that fell into a bin with width of 0.01.

## Acknowledgements

I thank Mike Ford, Phil Hedrick, Paul Moran, Robin Waples, and an anonymous reviewer for comments that improved this manuscript.

## References

Cavalli-Sforza LL, Edwards AWF (1967). Phylogenetic analysis: models and estimation procedures. *Evolution* **21**: 550–570.

Excoffier L, Novembre J, Schneider S (2000). SIMCOAL: a general coalescent program for the simulation of molecular data in interconnected populations with arbitrary demography. *J Hered* **91**: 506–508.

Foulley J, Hill WG (1999). On the precision of estimation of genetic distance. *Genet Selection Evol* **31**: 457–464.

Hedrick PW (1999). Perspective: highly variable loci and their interpretation in evolution and conservation. *Evolution* **53**: 313–318.

Hudson RR (1990). Gene genealogies and the coalescent process. *Oxford Surveys Evolution Biol* **7**: 1–44.

Nei M (1972). Genetic distance between populations. *Am Naturalist* **106**: 283–292.

Nei M (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* **89**: 583–590.

Nei M (1987). *Molecular Evolutionary Genetics*. Columbia University Press: New York.

Nei M, Tajima F, Tateno Y (1983). Accuracy of estimated phylogenetic trees from molecular data. *J Molec Evol* **19**: 153–170.

Sanghvi LD (1953). Comparison of genetical and morphological methods for a study of biological differences. *Am J Phys Anthropol* **11**: 385–404.

Takezaki N, Nei M (1996). Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. *Genetics* **144**: 389–399.

Weir BS, Cockerham CC (1984). Estimating *F*-Statistics for the analysis of population structure. *Evolution* **38**: 1358–1370.