Heredity (2005) 94, 33–36
© 2005 Nature Publishing Group All rights reserved 0018-067X/05 $30.00
www.nature.com/hdy

# Do polymorphic loci require large sample sizes to estimate genetic distances?

ST Kalinowski
*Department of Ecology, Montana State University, Bozeman, MT 59717, USA*

The coefficient of variation of estimates of three genetic distances (standard genetic distance of Nei, chord distance, $F_{ST}$) was examined with computer simulation to determine if large samples (per population) are necessary to precisely estimate genetic distances at loci with high levels of polymorphism. These simulations showed that loci with high mutation rates produce estimates of genetic distance with lower coefficients of variation than loci with lower mutation rates – without requiring larger sample sizes from each population. In addition, the rate at which increasing sample sizes decreases the coefficient of variation of estimates of genetic distances was shown to be approximately determined by the value of $F_{ST}$ between the populations being sampled. When $F_{ST}$ was greater than 0.05, sampling fewer than 20 individuals (per population) should be sufficient. When $F_{ST}$ was less than 0.01, sampling 100 individuals (per population) or more will be useful.

*Heredity* (2005) **94,** 33–36. doi:10.1038/sj.hdy.6800548
Published online 25 August 2004

**Keywords:** genetic distance; estimation; sampling; study design; mutation rate

## Introduction

Evolutionary and conservation geneticists frequently rely on neutral molecular data to describe population structure. In the past 30 years, a parade of molecular markers have been used, from blood proteins to microsatellites. This progression has been motivated, at least in part, by a search for loci with more variation. Loci with many alleles, such as microsatellite loci, have unprecedented ability to detect and describe genetic differences between populations (eg Hedrick, 1999; Kalinowski, 2002a).

However, loci with scores of alleles have forced population geneticists to re-evaluate how genotypic data are analyzed and interpreted. The most fundamental result of this re-evaluation has been increased awareness that statistically significant genetic differences are not always biologically or evolutionarily significant (eg Waples, 1998; Hedrick, 1999). One question that has received little attention in the literature is how high levels of polymorphism affect study design. This may be because most geneticists using highly polymorphic microsatellite loci have already concluded that large sample are needed to estimate genetic distances at loci with many alleles. All the literature that I am aware of supports this belief. For example, Nei (1978) analyzed the sampling variances of his genetic distances, and concluded that 'more individuals should be examined when heterozygosity is high than when it is low.' Baverstock and Moritz (1996) concluded that 'it is clear that large sample sizes are needed' to describe population structure at hypervariable loci. Most recently, Ruzzante (1998) used computer simulations to show that polymorphic loci have high sampling variances when sample size is small. Each of these authors, however, examined the relationship between samples sizes and sampling variance.

This focus on sampling variance has been misleading for two reasons. First, genetic distances are derived measures of genetic differentiation. They do not necessarily have high sampling variances when estimates of allele frequencies are imprecise. For example, high mutation rates usually decrease the sampling variance of $F_{ST}$. Second, sampling variances are not always an appropriate measure of precision to compare study design strategies. Consider the standard genetic distance, $D_S$, between two populations that have been isolated for $t$ generations. The sampling variance of $D_S$ will be higher at loci with high mutation rates than at loci with low mutation rates. However, the parametric genetic distance will also be higher

$$D_S = 2\mu t \qquad (1)$$

(where $\mu$ is the infinite alleles mutation rate) (Nei, 1972). This must be taken into consideration when comparing variances, and the coefficient of variation is a useful statistic to do this (see the Appendix for a mathematical discussion of why the coefficient of variation is a useful statistic for examining study design). The purpose of this paper is to explore the relationship between sample size, polymorphism, and the coefficient of variation of genetic distances.

## Methods

I examined the relationship between sample size, mutation rate, and the coefficient of variation of three popular genetic distances: the standard genetic distance of Nei, $D_S$, the chord distance of Nei (1983), $D_A$, and the Weir and Cockerham estimator of $F_{ST}(\theta)$, (1984) (see Excoffier (2001) for a discussion of the relationship

*Correspondence: ST Kalinowski, Department of Ecology, Montana State University, Bozeman, MT 59717, USA.*
*E-mail: skalinowski@montana.edu*

Received 1 December 2003; accepted 1 June 2004; published online 25 August 2004

between $\theta$ and $F_{ST}$). I chose these three genetic distances because they are commonly used and because they have substantially different evolutionary properties (see Kalinowski (2002b) for a review). I used computer simulation (see below) to see how increasing the sample size per population decreases the coefficients of variation of estimates of these genetic distances for loci with different levels of polymorphism. Because the expected amount of polymorphism at a locus is proportional to mutation rate ($\mu$) and effective population size ($N_e$), I explored how both $\mu$ and $N_e$ affect estimation of genetic distances.

I examined the sampling properties of $D_S$, $D_A$, and $F_{ST}$ in two simple evolutionary models: an isolation model of population divergence and an equilibrium model of migration. In the 'isolation' model, a randomly mating population of $N_e$ individuals is instantly divided into two populations that each has the same effective size as the ancestral population. The two populations formed by this fragmentation event remain completely isolated for $t$ generations (at which point sampling occurs). I included three population sizes ($N_e = 500$, 5000, and 50 000) and three divergence times ($t = 50$, 500, and 5000) in my simulations. In the equilibrium 'migration' model, two populations of equal and constant effective size ($N_e$) exchange migrants at a rate of $m$. I included three population sizes ($N_e = 500$, 5000, and 50 000) and three migration rates ($m = 0.01$, 0.001, 0.0001) in my simulations.

In both evolutionary models, I simulated data for loci with four different mutation rates ($10^{-6}$, $10^{-5}$, $10^{-4}$, $10^{-3}$). All mutations were unique (infinite alleles mutation). These mutation rates (and the effective population sizes listed above) produced gene diversities within populations that ranged from approximately 0.002 to greater than 0.995. Sample size was varied from two individuals per population to 256 individuals per population. The number of loci in simulated data sets was varied from 2 to 256.

The amount of population differentiation in these models can be measured by $F_{ST}$. Most formulations of $F_{ST}$ are a function of mutation rates. However, formulae based solely on demographic variables can be obtained by taking the limit of the mutation rate as it goes to zero (Slatkin, 1991). Weir and Cockerham's estimator of $F_{ST}$ is then equal to

$$F_{ST} = \frac{t}{t + 2N}$$

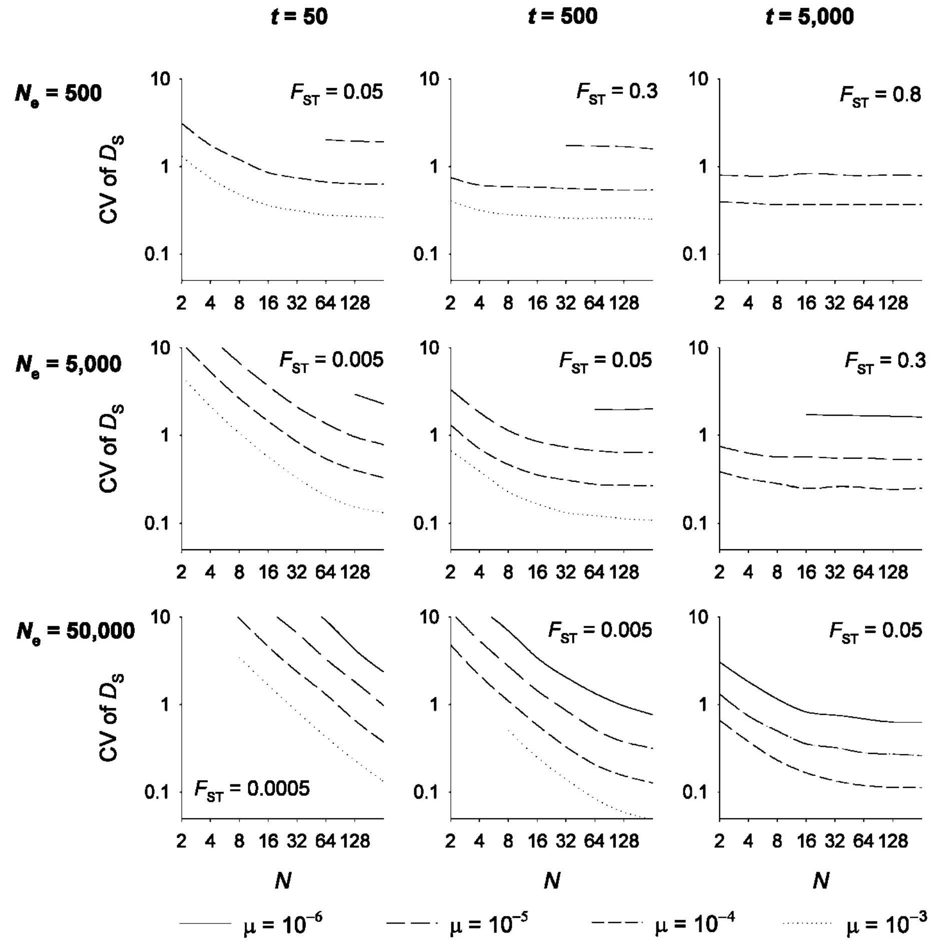

**Figure 1** The influence of mutation rate ($\mu$) and sample size (per population) on the coefficient of variation (CV) of estimates of $D_S$ (Nei, 1978) in an isolation model of population divergence. In this model, a population of $N_e$ individuals is instantly and permanently split into two completely isolated populations. Sampling occurs $t$ generations after population fragmentation. All samples have 16 loci. The parametric value of $F_{ST}$ for the populations is shown in each graph. Results for $D_A$ and $F_{ST}$ are qualitatively indistinguishable (not shown).

for the isolation model, and

$$F_{ST} = \frac{1}{1 + 8Nm}$$

for the migration model (Slatkin, 1991; Weir and Cockerham, 1984; Excoffier, 2001).

Genotypic data were simulated using the coalescent approach (eg, Hudson, 1990; Felsenstein, 2003) with a computer program that I wrote for this purpose. The method of Ford (1998) was used for the isolation model.

The coefficient of variation of $D_S$, $D_A$, and $F_{ST}$ was estimated by calculating the standard deviation and average value of 1000 simulated estimates of $D_S$, $D_A$, and $F_{ST}$. Calculations were performed in Microsoft Access.

Estimates of $D_S$ are not defined when two samples have no alleles in common, and estimates of $F_{ST}$ are not defined when all loci are fixed for the same allele. Therefore, such samples were removed from the analysis. The frequency of these excluded samples was recorded. If over 10% of the samples were excluded, then no results are reported for the combination of effective population size, mutation rate, sample size, that produced the undefined estimates.

## Results and discussion

All three genetic distances ($D_S$, $D_A$, and $F_{ST}$) displayed remarkably similar statistical properties, so I present representative results (Figures 1 and 2). Four trends were observed: three well known, one novel. First, large samples had a lower coefficient of variation than small samples. Second, increasing the sample size (per population) produced diminishing returns: at some point, sampling more individuals had little effect upon the coefficient of variation of the genetic distance. Third, loci with high mutation rates produced lower coefficients of variation than loci with low mutation rates.

What was interesting, was that the rate at which increasing sample size decreased the coefficient of variation was determined solely by the amount of differentiation between the populations – and not the mutation rate or the amount of variation at the loci. This was true for both the isolation and the migration models. More individuals should be sampled when the amount of differentiation is small than when it is large.

$F_{ST}$ proved to be a convenient measure of population differentiation for study design. Figures 1 and 2 suggest that 20 individuals is a reasonable maximum sample size when the parametric value of $F_{ST}$ is 0.05 and 100 individuals is a reasonable maximum sample size when

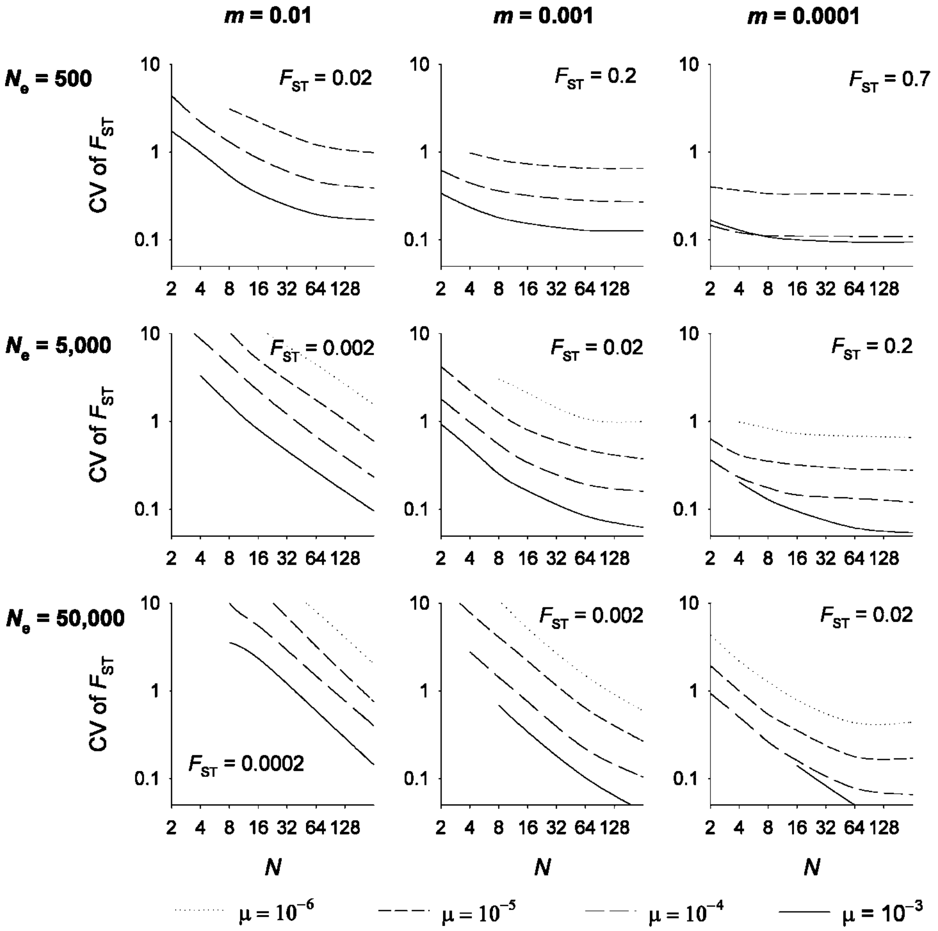

**Figure 2** Influence of mutation rate ($\mu$) and sample size (per population) on the coefficient of variation (CV) of estimates of $F_{ST}$ in an equilibrium migration model. In this model, two populations of $N_e$ individuals exchange migrants at a rate of $m$. All samples have 16 loci. The parametric value of $F_{ST}$ for the populations is shown in each graph. Results for $D_S$ and $D_A$ are qualitatively indistinguishable.

the parametric value of $F_{ST}$ is equal to 0.01. One implication of these results is that there is more benefit to collecting large samples from large populations than from small populations. This is because, all else being equal, $F_{ST}$ between large populations is smaller than $F_{ST}$ between small populations.

These results extend the seminal work of Nei and collaborators (see Nei, 1987 for a review). Nei and collaborators showed that sample sizes can be small when divergence times are large, but did not examine how effective population size or mutation rate affects the sampling properties of genetic distances. Foulley and Hill (1999) recently showed that only a few individuals need to be sampled to estimate the Sanghvi genetic distance when divergence times are large, but did not relate this genetic distance to effective population size or mutation rate.

Increasing the number of individuals in a study is not the only way to decrease the coefficient of variation of estimates of genetic distance. Increasing the number of loci will also improve the precision of estimates of genetic distance (see Nei, 1987 for a review). In fact, when population differentiation is substantial (eg $F_{ST} > 0.2$), increasing the number of loci is the only method for improving estimates of genetic distances. However, if enough loci are available, reliable estimates of genetic distances can be obtained from very few individuals. Figures 1 and 2 depict results for 16 loci. The shape of the curves in these figures, however, was independent of the number of loci examined. If fewer than 16 loci are sampled, the lines in the figures are shifted upwards (higher coefficient of variation). If more than 16 loci are sampled, the lines in the figures are shifted downwards (lower coefficient of variation).

## References

Excoffier L (2001). Analysis of population subdivision. In: Balding DJ, Bishop M, Cannings C (eds) *Handbook of Statistical Genetics*, John Wiley & Sons, Ltd: New York. pp 271–307.

Felsenstein J (2003). *Inferring Phylogenies*, Sinauer Associates: Sunderland, MA.

Ford M (1998). Testing models of migration and isolation among populations of Chinook salmon (*Oncorhynchus tschawytscha*). *Evolution* **52**: 539–557.

Foulley J, Hill WG (1999). On the precision of estimation of genetic distance. *Genetics, 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 Surv Evol Biol* **7**: 1–44.

Kalinowski ST (2002a). How many alleles should be used to estimate genetic distances? *Heredity* **88**: 62–65.

Kalinowski ST (2002b). Statistical properties of three genetic distances. *Mol Ecol* **11**: 1263–1273.

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, NY.

Ruzzante DE 1998. A comparison of several measures of genetic distance and population structure with microsatellite data: bias and sampling variance. *Can J Fish Aquat Sci* **55**: 1–14.

Slatkin M (1991). Inbreeding coefficients and coalescent times. *Genet Res* **58**: 167–175.

Waples RS (1998). Separating the wheat from the chaff: patterns of genetic differentiation in high gene flow species. *J Hered* **89**: 438–450.

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

## Appendix

Consider three populations, *A*, *B*, and *C*, and the genetic distances between them, $d_{AB}$ and $d_{BC}$. In many applications, the magnitude of $d_{AB}$ and $d_{BC}$ is less informative than the ratio

$$r = \frac{d_{AB}}{d_{BC}} \tag{A1}$$

For example, consider a phenogram constructed from Nei's standard genetic distance. The length of the branches is not particularly informative, because that length is determined by the mutation rate, and this is usually unknown. What is informative, is the relative lengths of each branch. Here, I examine how sampling error affects estimates of *r*.

Let $\hat{d}_{AB}$ and $\hat{d}_{BC}$ represent estimates of $d_{AB}$ and $d_{BC}$, respectively. These estimates can be modeled

$$\begin{aligned} \hat{d}_{AB} &= d_{AB} + \varepsilon_{AB} \\ \hat{d}_{BC} &= d_{BC} + \varepsilon_{BC} \end{aligned} \tag{A2}$$

where $\varepsilon_{AB}$ and $\varepsilon_{BC}$ are random variables that model sampling error. For simplicity, assume the expected value of $\varepsilon_{AB}$ and $\varepsilon_{BC}$ is zero (ie $\hat{d}_{AB}$ and $\hat{d}_{BC}$ are unbiased estimators). In this model, the standard errors (*SD*) of $\hat{d}_{AB}$ and $\hat{d}_{BC}$ are equal to the standard deviations of $\varepsilon_{AB}$ and $\varepsilon_{BC}$

$$\begin{aligned} SD\left(\hat{d}_{AB}\right) &= SD(\varepsilon_{AB}) \\ SD\left(\hat{d}_{BC}\right) &= SD(\varepsilon_{BC}) \end{aligned} \tag{A3}$$

The estimate of *r*, $\hat{r}$, is modeled

$$\hat{r} = \frac{d_{AB} + \varepsilon_{AB}}{d_{BC} + \varepsilon_{BC}} \tag{A4}$$

The goal of study design is to determine how many loci and/or individuals must be sampled so that $\hat{r}$ is likely to approximate *r*. The approximation will be good when

$$\begin{aligned} |\varepsilon_{AB}| &\ll d_{AB} \\ |\varepsilon_{BC}| &\ll d_{BC} \end{aligned} \tag{A5}$$

This, in turn, is most likely to occur when

$$\begin{aligned} SD(\varepsilon 183_{AB}) &\ll 183 d_{AB} \\ 183 SD(\varepsilon_{BC}) &\ll 183 d_{BC} \end{aligned} \tag{A6}$$

Substitution and rearrangement show that $\hat{r}$ is likely to approximate *r* when

$$\begin{aligned} &\frac{SD\left(\hat{d}_{AB}\right)}{d_{AB}} \ll 1 \\ &\frac{SD\left(\hat{d}_{BC}\right)}{d_{BC}} \ll 1 \end{aligned} \tag{A7}$$

Note that the left-hand terms in equations (A7) are the coefficients of variation of $\hat{d}_{AB}$ and $\hat{d}_{BC}$.