Molecular Ecology (2002) **11**, 1263–1273

INVITED REVIEW

Evolutionary and statistical properties of three genetic distances

STEVEN T. KALINOWSKI

Conservation Biology Division, Northwest Fisheries Science Center, National Marine Fisheries Service, 2725 Montlake Boulevard East, Seattle, WA 98112, USA

Abstract

Many genetic distances have been developed to summarize allele frequency differences between populations. I review the evolutionary and statistical properties of three popular genetic distances: D_S, D_A, and θ, using computer simulation of two simple evolutionary histories: an isolation model of population divergence and an equilibrium migration model. The effect of effective population size, mutation rate, and mutation mechanism upon the parametric value between pairs of populations in these models explored, and the unique properties of each distance are described. The effect of these evolutionary parameters on study design is also investigated and similar results are found for each genetic distance in each model of evolution: large sample sizes are warranted when populations are relatively genetically similar; and loci with more alleles produce better estimates of genetic distance.

Keywords: effective size, estimation, genetic distance, isolation, migration, mutation

Received 2 November 2001; revision received 10 April 2002; accepted 10 April 2002

Introduction

Analysis of genotypic data from neutral loci is an important method for describing the patterns of genetic variation within species and inferring the evolutionary processes that give rise to those patterns. Genotypic data are notoriously multivariate: the frequency of each allele at each locus is usually different in each population. Genetic distances are metrics that summarize these differences in an overall measure of differentiation for a pair of populations. Generally, a matrix of pair wise genetic distances between a set of populations is estimated. This matrix is then often visualized with phenograms, isolation by distance plots, principal component analysis, or multidimensional scaling plots.

Many genetic distances have been developed, of which a few remain in regular use (see Nei 1987 for a review of several genetic distances). Each of these genetic distances has unique evolutionary and statistical properties, and evolutionary relationships inferred from each genetic distances can be quite different. For example, King *et al*. (unpublished) observed that the standard genetic distance of Nei showed North American and European populations of Atlantic salmon (*Salmo salar*) to be much more distinct than the original chord distance. Surprisingly, there has been insufficient examination of these genetic distances to confidently explain the reason behind differences such as this.

Analysis of highly polymorphic microsatellite loci has provided population geneticists with new opportunities and new challenges. For example, microsatellite data have unprecedented power to detect and describe small genetic differences between populations. Apparently, this is because microsatellite loci have a much higher mutation rate than allozyme or mitochondrial loci that have previously been the mainstay of population structure studies. The high mutation rate and unusual mutation mechanism of microsatellite loci, however, have also forced population geneticists to reconsider how genotypic data should be analysed and interpreted. One of the most substantial results from this discussion has been increased recognition that there is an important distinction between statistically significant genetic differences between populations and evolutionarily or biologically significant differences (e.g. Waples 1991; Hedrick 1999). The purpose of genetic data

Correspondence: Steven Kalinowski. Fax: 206 860–3335; E-mail: Steven.Kalinowski@noaa.gov

therefore are usually not to demonstrate that two populations are different, but to reveal how different they are (see Anderson *et al*. 2000 for a discussion of the difference between hypothesis tests and estimation). Therefore, selecting a genetic distance appropriate for the specific study being performed and estimating it accurately is important.

Most of the recent work on this subject has focused on identifying a genetic distance appropriate for analysing microsatellite loci. Most mutations at microsatellite loci add or subtract one repeat motif and several sized based genetic distances have been developed. These mutation explicit distances are commonly used, but they have not replaced 'traditional' genetic distances, i.e. genetic distances not specifically developed for microsatellite loci. A survey of the papers published in *Molecular Ecology* during the year 2000 (S. Kalinowski unpublished) shows that traditional genetic distances were used by the majority of authors, and I restrict my discussion to three of these genetic distances.

The purpose of this paper is to describe the evolutionary and statistical properties of three popular traditional genetic distances that must be recognized in order to efficiently design population structure studies and interpret the genetic data obtained. I begin by examining how evolutionary parameters such as divergence time, migration rate, effective population size, and mutation rate affect the parametric value of these genetic distances. Then I consider how these evolutionary factors influence estimation of genetic distances. Throughout this investigation, I have taken a qualitative approach and rely heavily on graphs to illustrate basic principles. I examine a wide range of evolutionary parameters, but emphasize the properties of loci with high mutation rates in relatively similar populations. Many of the properties of genetic distances that I describe have previously been discussed in the literature (see Nei 1987; Chakraborty & Rao 1991; Weir 1996; Nei & Kumar 2000; for reviews). However, this paper is not intended to be a review of the genetic distance literature. Instead, this paper attempts to provide a comprehensive and consistent comparison of three genetic distances that population geneticists frequently use.

General properties of D_S, D_A, and θ

I chose to evaluate three genetic distances: the standard genetic distance of Nei (1972, 1978), D_S, the chord distance of Nei *et al*. (1983), D_A, and the Weir & Cockerham (1984) analogue of F_{ST}, θ. These three genetic distances were chosen from among the many available genetic distances because they are all relatively popular, and because they have distinct properties.

The standard genetic distance of Nei (1972, 1978) remains one of the most commonly used genetic distances. For populations X and Y with r loci and m alleles per locus, the standard distance is defined as

$$D_S = -\ln\left(J_{XY}/\sqrt{J_{XX}J_{YY}}\right)$$

where $J_{XY} = \sum_{i=1}^{m}\sum_{j=1}^{r} x_{ij}y_{ij}/r, J_{XX} = \sum_{i=1}^{m}\sum_{j=1}^{r} x_{ij}^2/r, J_{YY} = \sum_{i=1}^{m}\sum_{j=1}^{r} y_{ij}^2/r$, x_{ij} is the frequency of the i^{th} allele at the j^{th} locus in population X, and y_{ij} is the frequency of the i^{th} allele at the j^{th} locus in population Y. The parametric value of D_S between two populations that became separated t generations in the past is approximately

$$D_S \approx 2\mu t \quad (1)$$

where μ is the infinite alleles mutation rate at the loci examined. This expression assumes that fragmentation of the ancestral population was instantaneous and complete, and that each population has had a constant effective size equal to the effective size of the original ancestral population. Note that D_S increases linearly with time from zero to infinity and will have a value proportional to the mutation rate. A formula is available for obtaining nearly unbiased estimates of D_S from genotypic data (Nei 1978).

The D_A distance of Nei (Nei *et al*. 1983) is a modification of the original Cavalli-Sforza chord distance (1967)

$$D_A = 1 - \sum_{i=1}^{m}\sum_{f=1}^{r}\frac{\sqrt{x_{ij}y_{ij}}}{r}.$$

Its maximum value of 1.0 is achieved when two populations share no alleles at any loci. The D_A distance has proven to be useful for reconstructing phylogenies (Takezaki & Nei 1996). There currently is no method for obtaining unbiased estimates of D_A.

Wright's F_{ST} is one of the most fundamental measures of population structure available. Several analogous measures (e.g. θ, β, and G_{ST}) have been developed to describe differentiation between populations (see Excoffier 2001 for a review). These measures have different mathematical foundations (Nei & Kumar 2000; Excoffier 2001, and references within) and represent distinct but related concepts. The relative merits of each measure have not been resolved, but, in practice, estimates of these statistics are generally similar. I examine θ (See Weir 1996 for formulae) because it is most commonly used. In this investigation, I use θ as a genetic distance between two populations instead of a fixation index among many populations. If two populations are completely isolated for short period of time and the effective population size of each population is equal and constant, then θ will roughly equal $t/2N_e$ (See Nei 1987 for a discussion of the relationship between F_{ST} and divergence time). One important characteristic of θ is that its maximum value will only be 1.0 when populations are fixed for alternative alleles. If there is polymorphism present in populations, the maximum value — obtained

when populations do not share alleles — will be less than 1.0. This maximum value appears to be the homozygosity present within the populations being compared. Formulae are available to obtain virtually unbiased estimates of θ (Weir & Cockerham 1984; Weir 1996).

Evolutionary models

I have examined the behaviour of D_S, D_A, and θ in two simple evolutionary models: an isolation model of population divergence and an equilibrium migration model. In the 'isolation' model, a randomly mating population of N_e individuals is instantly divided into two populations that each have the same effective size as the ancestral population. The populations remain completely isolated from each other for t generations. In the 'migration' model, two populations of equal and constant effective size (N_e) exchange migrants at a rate of m (where m indicates the migration rate into each population).

Simulation method

Genetic distances can be examined empirically, experimentally, analytically, or through computer simulation. Each approach is useful, but only the latter approach (computer simulation) can address all of the questions that I will address. For example, no formula is available for the coefficient of variation for estimates of D_A between two populations separated for 1000 generations at loci with a mutation rate of 10^{-4}. The advantage of computer simulation is that the precise details of the evolutionary history giving rise to the simulated data is known. This contrasts with most empirical data. The disadvantage of using simulated data to study evolution is that the realism of the simulated evolutionary processes is difficult to evaluate. This is not a problem for this investigation, because my goal is to explore how different evolutionary processes affect genetic distances.

I used coalescent simulation to estimate the parametric genetic distance between populations. To do this I average simulated data containing 100 loci and 500–5000 individuals (I used 500 individuals when simulating data from populations having an effective size of 500 and 5000 individuals when populations were larger). The coalescent approach that I used is described by Hudson (1990). In the isolation model that I used, genes coalesce within their respective populations for the first t generations in the past. When over t generations has passed, genes from both populations are pooled. In the migration model that I used, the timing of coalescent and migration events and the relative probabilities of each is given by Hudson (1990; page 20).

I examined two models of mutation: infinite alleles mutation (IAM), and single stepwise mutation (SSM). The IAM model assumes that each mutation creates a new and unique allele and that there is no limit to the number of alleles possible at a locus. The SSM model assumes that mutation either adds or subtracts a repeat motif from an allele. I have assumed that each event is equally likely and that there are no bounds or restrictions to the number of repeat units possible at a locus.

The effect of evolutionary parameters upon genetic distances

Inferring the evolutionary history of populations from genetic data is difficult because genetic differences observed between populations can be explained by an infinite number of evolutionary histories. For example, a small value of D_S between two populations might indicate that gene flow is relatively common between them, or that there is no gene flow between the populations at present, but that they recently were part of a larger population. Other variables besides m (rate of gene flow) and t (length of population isolation) also affect the genetic distance between populations. Effective population size, mutation rate, and mutation mechanism (e.g. IAM vs. SSM) also play important roles in determining the genetic distance between populations. therefore, a thorough understanding the effect of these evolutionary variables upon genetic distances is necessary to evaluate genetic data.

Isolation and migration

I will begin by examining how D_S, D_A, and θ are affected by isolation time and migration rate in a simple example (N_e = 5000, μ = 10^{-4} IAM). In the isolation model (Fig. 1a), each of the three genetic distances increases approximately linear with time for at least a couple of thousand of generations. as expected, D_S is precisely linear with time.

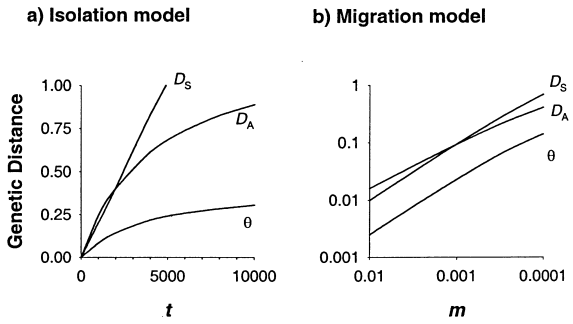

Fig. 1 The effect of duration of population isolation (a) and migration rate (b) upon the parametric value of three genetic distances: the standard genetic distance of Nei (1972), D_S, the chord distance as formulated by Nei *et al.* (1983), D_A, the Weir & Cockerham (1984) analogue of F_{ST}, θ, and linearized θ, –ln(1 – θ). The effective size of each population is 5000 (constant through time), and the mutation rate is 10^{-4} (IAM).

After a few thousand generations, the other two genetic distances noticeably approach a maximum value. Figure 1(a) shows that θ approaches 0.33 as the populations become maximally differentiated (i.e. they share no alleles). This value is the homozygosity within each population. In the migration model, each of the genetic distances is inversely related to the migration rate.

These results only apply to the parametric genetic distance between populations. Estimates of these genetic distances based on a small number of loci or individuals will depart from these trends.

Effective population size

The effective size of populations, N_e, is one of the most fundamental parameters in population genetics. The effective size of populations determines how much genetic variation can be maintained in populations that are in mutation-drift equilibrium, and determines how quickly allele frequencies change with genetic drift.

I examined how effective population size affects the genetic distance between populations by comparing the genetic distance between pairs of populations that have an effective size of either 500, 5000, or 5000. I assume, for now, that the effective size is constant through time and that the mutation rate is 10^{-4} (IAM). Each of the genetic distances behaves similarly in both evolutionary models (Fig. 2). D_S is unaffected by effective population size; D_A is virtually unaffected by population size. In contrast, and as expected, θ is strongly affected by effective population size. θ is large when populations are small, and small when populations are large. An interesting effect of N_e upon θ observable in the isolation model is that the maximum value of θ is greater for small populations than large populations. This is because, at equilibrium, small populations have less genetic diversity than large populations, and as mentioned above, the maximum value of θ appears to be the homozygosity within the populations.

Mutation rate

The effect of mutation rate upon genetic distances may be important to consider when comparing genetic distances calculated from different loci. The mutation rate of loci is seldom known, but for a given pair of populations, it is expected to be proportional to the amount of polymorphism observed.

In the isolation model of evolution, D_S and D_A increase more quickly with time at loci with high mutation rates than at loci with low mutation rates (Fig. 3a). In contrast, θ is much less affected by mutation rate. For example, the expected value of θ for loci with a mutation rate of 10^{-5} is virtually the same as for loci with a mutation rate of 10^{-6} (at least for the combinations of parameters examined in Fig. 3). The principal effect of mutation rate upon the parametric value of θ is that the mutation rate helps determine the maximum value that θ can obtain. The mutation rate also affects how quickly genetic distances approach their maximum value (if they have one). When the mutation rate is very high (e.g. 10^{-3}) D_A approaches its maximum value of 1.0 quickly. θ also approaches its maximum value quickly for loci with such high mutation rates. Notice, however, that this maximum value is quite small when mutation rates are high. These genetic distances behave fairly similarly in the migration model: D_S and D_A are high for loci with mutation rates; θ is almost unaffected by mutation rate unless the mutation rate is exceptionally high.

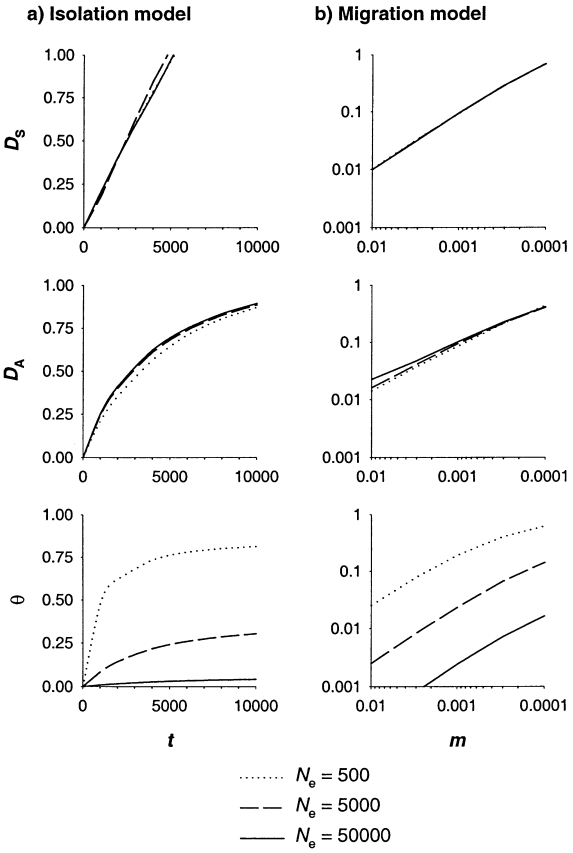

Fig. 2 The effect of population size upon the parametric value of three genetic distances (D_S, D_A, θ) between two populations isolated for t generations (a) or connected by a migration rate of m (b). Population size, N_e, varies for each pair of populations, but is constant through time. The mutation rate is 10^{-4} (IAM).

Mutation mechanism

Most mutations at microsatellite loci are believed to add or subtract a single repeat motif to the locus (see Goldstein & Schlotterer 1999 for reviews). This stepwise mutation model makes backwards mutation and homoplasy common. This contrasts with the IAM of mutation, in which each

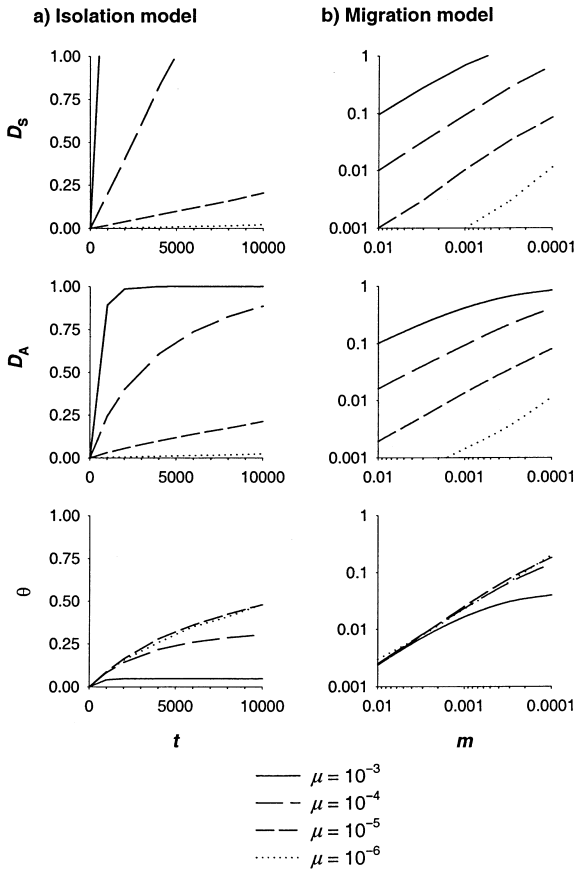

Fig. 3 The effect of mutation rate (IAM) upon the parametric value of three genetic distances (D_S, D_A, θ) between two populations isolated for t generations (a) or connected by a migration rate of m (b). Effective population size is 5000 and constant.

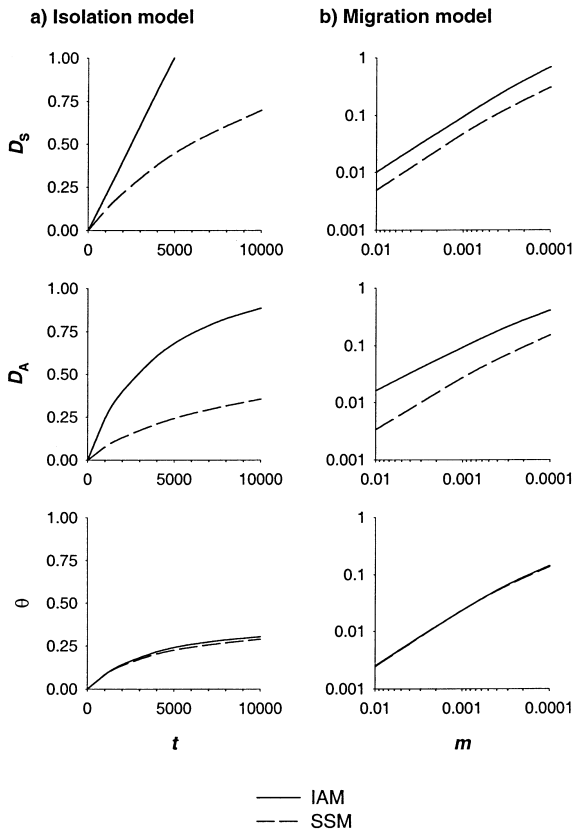

Fig. 4 The effect of mutation model upon the parametric value of three genetic distances (D_S, D_A, θ) between two populations isolated for t generations (a) or connected by a migration rate of m (b). The mutation rate is 10^{-4} in all cases, and an effective size of 5000 is assumed.

mutation results in an unique allele. The most important consequence of stepwise mutation for the isolation model is that genetic distances derived to increase linearly with time (e.g. D_S) will not do so (Fig. 4a). This nonlinearity however, is not necessarily severe. Consider the genetic distance between two populations having a constant population size of 5000 individuals (Fig. 4a) and a mutation rate of 10^{-4}. Over short periods of time, $t < 1000$ generations, IAM and SSM mutation produce very similar genetic distances (note: this is not readily apparent in Fig. 4a, but re-scaling the figure would show this). Over longer periods of time, $t > 1000$ generations, SSM results in a substantially smaller genetic distance for D_S and D_A. This decreased the linearity of D_S, but actually appeared to increase the linearity of D_A by slowing its approach to its maximum value of 1.0. In contrast, the parametric value of θ was virtually unaffected by the mutational mechanism. The migration model shows similar properties: SSM mutation leads to lower values of D_S and D_A than IAM, but does not effect θ.

Population size reduction

In all of the evolutionary models considered so far, effective population size has been assumed constant. In natural populations, population size will change. Therefore, examining the consequences of a change in population size is useful. Effective population size could change in many possible ways, but I will examine only one scenario: an instantaneous and permanent reduction in size. This special case is particularly significant, because it has the potential of rapidly increasing the genetic distance between populations. In the isolation model, I assume that the reduction in population size occurs at the time the ancestral population fragments into two populations. In the migration model, I assume that the migration rate between the two populations remains the same.

The three genetic distances being examined here behave quite differently in the isolation and migration models (Fig. 5). In the isolation model, all of the genetic distances increase more rapidly when the population decreases to a small size than when population size is constant (Fig. 5a). Over the length of time examined for this model (1000

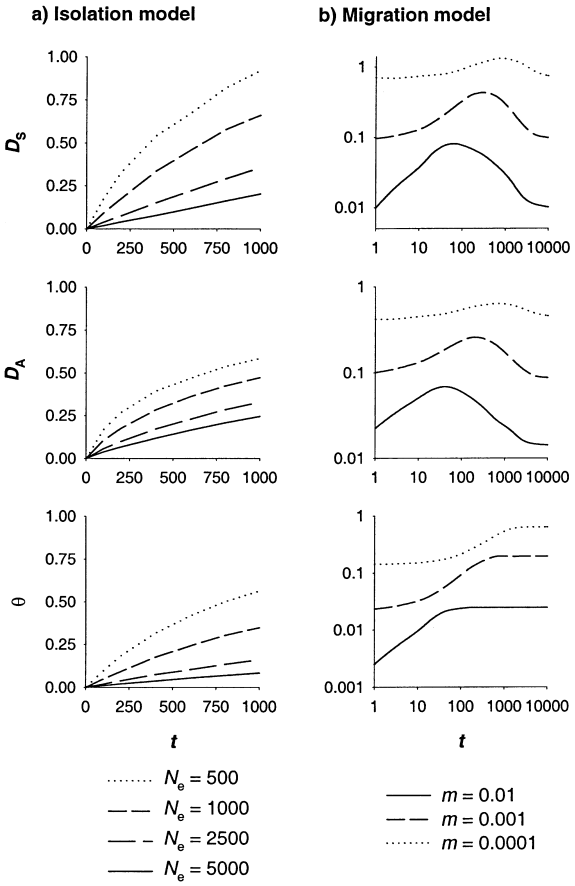

Fig. 5 The effect of population size reduction upon the parametric value of three genetic distances (D_S, D_A, θ) for two populations isolated for t generations (a) or connected by a migration rate of m (b). In each evolutionary model the ancestral population had an effective size of 5000 individuals. In the isolation model (a), the population fragments had a constant population size of either 2500, 1000, or 500. In the migration model (b), each population had a constant size of 500 after time 0. The mutation rate is 10^{-4} (IAM).

generations), D_A, shows the least linear response. In the migration model, all of the genetic distances initially increase in value. D_S and D_A then very slowly decline until they approach their original (prebottleneck) value. The slowness of this approach to equilibrium is noteworthy; thousands of generations are required. In contrast, θ quickly approaches a new and larger equilibrium values. A likely explanation for why D_S and D_A are slow to approach their equilibrium values is that they are affected by the amount of polymorphism within populations. This declines when the population size decreases, but takes a long time to reach its new equilibrium value. θ is less affected by the amount of polymorphism within populations, and therefore approach its equilibrium values quickly. Chakraborty & Nei (1977) provide an analytic examination of the effect of a short bottleneck upon the genetic distance between two populations.

Estimating genetic distances

Understanding the sampling properties of genetic distances is necessary for efficiently designing population structure studies and for accurately interpreting estimates of genetic distances. Modern laboratory technology enables population geneticists to gather unprecedented amounts of genetic data, but reliable estimation of genetic distances between populations often remains challenging because increasingly complex and subtle evolutionary questions are being asked. Therefore, efficient study design is necessary to prevent genetic data from being too imprecise to answer the questions asked of it. Understanding the sampling properties of genetic distances such as sampling variance and bias is important because it assists data interpretation.

Estimates of genetic distances should have two statistical properties: low sampling variance and minimal bias. Sampling variance is a measure of how much estimates of a parameter are likely to differ from the parameter being estimated, so a small variance is preferable to a large variance. Comparing the sampling variances of D_S, D_A, and θ, however, is not informative because each genetic distance has a different parametric value. Therefore, the coefficient of variation of estimates of each genetic distance is a more useful measure of sampling error than the sampling variance. This is the measure of variability that I will examine. Along with low sampling variance, estimates of genetic distances should have low sampling bias. Bias is the difference between the expected value of an estimate of a statistic and the actual value of the parameter being estimated.

Coefficient of variation

Informed study design for estimating genetic distances involves efficiently minimizing the coefficient of variation of estimates of genetic distances (see Appendix II). This involves deciding how many individuals to sample and how many loci to characterize. The impact of effective population size, length of population divergence or migration rate, and mutation rate on these decisions are important to consider.

Each of the genetic distances examined here respond to increasing sample sizes in a similar manner in both evolutionary models (Fig. 6). In all cases, increasing the sample size decreases the coefficient of variation. However, for all pairs of distinct populations, increasing sample size has a diminishing effect upon the coefficient of variation. At some point, increasing the sample size will have no discernable effect upon the coefficient of variation. The point at which increasing sample size brings diminishing returns is determined by the level of differentiation between the populations: when F_{ST} is small, large sample sizes are useful for reducing the coefficient of variation; when F_{ST} is large, large sample sizes are not useful for reducing the coefficient of variation (Fig. 6).

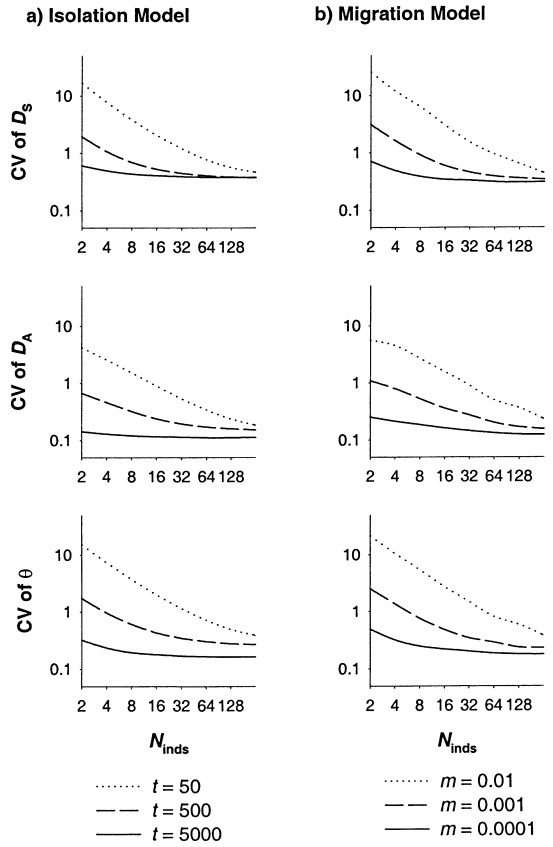

Fig. 6 The effect of sample size (N_{inds}) upon the coefficient of variation of estimates of 3 genetic distances (D_S, D_A, θ) based on 8 loci for populations of 5000 individuals separated for t = 50, 500, or 5000 generations or connected by a migration rate of m = 0.01, 0.001, or 0.0001 (b). The mutation rate is 10^{-4} (iam). Notice use of a log scale on all axes.

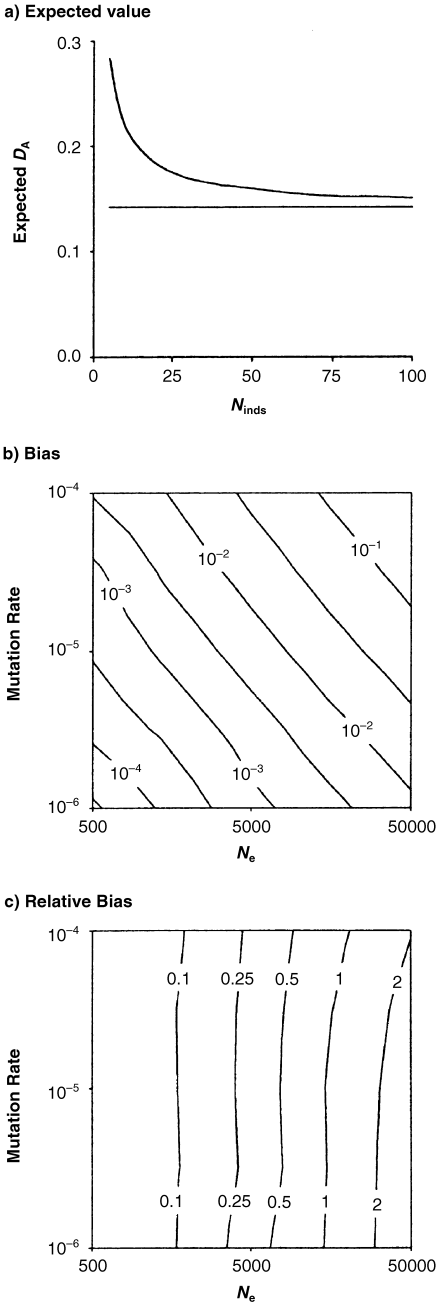

Fig. 7 The effect of sample size, mutation rate, and effective population size upon bias in estimates of D_A for two populations isolated for 500 generations. in graph a, the mutation rate is 10^{-4} (IAM) and the effective population size is 5000 individuals. The curve shows the expected value of estimates of D_A, and the line shows the parametric value being estimated. In graphs b & c, the sample size is 32 individuals. relative bias is defined here as bias divided by the parametric value being estimated.

The relationship between genetic polymorphism and the quality of estimates of genetic distances has received a lot of informal discussion. One question that is commonly asked is whether a few loci with many alleles produce better estimates of genetic distances than many loci with a few alleles. The answer appears to be that both approaches work equally well, as long as the amount of divergence between populations is not large (Kalinowski 2002). If the amount of divergence is low to moderate, the total number of independent alleles present at the loci examined appear to be a good indicator of the coefficient of variation of estimates of genetic distances. The number of independent alleles present at a locus is one less than the number of alleles observed at that locus. This is true for both IAM and SSM mutation. Again, each of the genetic distances appears to behave similarly (see Foulley & Hill 1999 for an analytic demonstration that the Sanghvi genetic distance shares these properties). If the amount of divergence is high, then sampling a large number of loci with a few alleles each will produce better estimates of genetic distance than a few loci with large number of alleles.

The discussion above has focused on the relationship between coefficient of variation and study design. Fortunately,

all three genetic distances behave very similarly in this regard. However, this does not mean that the coefficient of variation of each of these genetic distances are similar. A thorough examination of the coefficient of variation for each of these genetic distances has not been performed. However, all available evidence suggests that D_S has the highest coefficient of variation of the three genetic distances considered here, and that D_A has the lowest (Nei 1987; Kalinowski 2002; present study Fig. 7). This appears to be because D_S has a high interlocus variance. Apparently, this is the price of not having a maximum value. This may be why D_A is more successful at estimating the topology of phylogenetic trees than D_S (Takezaki & Nei 1996) or θ.

Bias

Formulae for D_S and θ have been developed that minimize the effect of sampling a limited number of individuals from populations. Both statistics are calculated as ratios, and unbiased estimates of the numerators and denominators of these ratios have been developed. Estimates of the ratio themselves are not unbiased. However, the amount of bias for estimates of D_S and θ are negligible as long as the amount of data collected (number of individuals sampled and number of loci examined) is not very small (Chakraborty & Rao 1991; Kalinowski unpublished), and I shall not consider D_S and θ here.

In contrast to D_S and θ, there is no unbiased estimators for D_A. Using the allele frequencies present in samples to estimate D_A produces inflated estimates when sample size is small (Fig. 7a). The magnitude of this bias appears to be proportional to the amount of variation present at loci (Fig. 7b). Therefore, bias is largest for loci with high mutation rates and in large populations (Fig. 7b). However, the relative bias, i.e. the amount of bias relative to the parametric genetic distance, is nearly independent of mutation rate for this genetic distances (Fig. 7c). This is because loci with a high mutation rate have a higher parametric distance. This leads to the conclusion that if sample sizes are being selected to minimize relative bias, loci with a high mutation rate do not require larger sample sizes than loci with a lower mutation rate, but large populations require larger samples than small populations. This result is particularly fortuitous because this result is completely concordant with the sampling recommendations developed to minimize the coefficient of variation discussed above (large samples are warranted when F_{ST} is relatively small, which occurs when N_e is large).

Conclusions

I have emphasized two points: (i) each of these three genetic distances has unique evolutionary properties, and (ii) each genetic distance has relatively similar sampling properties.

Selecting the most appropriate genetic distance for an investigation requires careful consideration of the most likely evolutionary history of the populations involved and the specific goals of the investigation. None of the genetic distances discussed here stand out as best in all circumstances. However, each genetic distance should work well in most circumstances, if the idiosyncrasies of the distance are recognized.

A few comments regarding microsatellite loci and sized based genetic distances are warranted. The principal reason for developing sized based genetic distances for microsatellite loci is to obtain a distance measure that increases linearly with time. Mutation mechanism, however, is only one of several evolutionary variables that affect the linearity of genetic distances. Furthermore, there is little reason to suspect that departure from the infinite alleles model of mutation has a larger effect upon the linearity of genetic distances than the other processes described above. For example, variation in population size across time and space could easily have a much stronger effect upon genetic distances than mutation mechanism. Therefore, I suspect that the potential benefit of accounting for stepwise mutations will often be small, especially when the length of population isolation has been small. On the other hand, two potential disadvantages of sized-based genetic distances are possible, and they may be serious. First, microsatellite loci appear to mutate in complex ways (see Goldstein & Schlotterer 1999 for reviews). For example, occasional mutations adding or subtracting many repeat units can have a strong effect upon size-based distances. Second, sized-based genetic distances appear to inherently have a high sampling variance (see Goldstein & Pollock 1997 for a review of these issues).

Whichever genetic distance is used to summarize the genetic differences between populations, the greatest challenge will be deciding what evolutionary processes produced the observed pattern, and evaluating what biological significance that has for the populations.

Acknowledgements

I would like to thank P. Hedrick, J. Johnson, P. Moran, M. Nei, and Robin Waples for comments that improved this manuscript.

References

Anderson DR, Burnham KP, Thompson WL (2000) Null hypothesis testing: problems, prevalence, and an alternative. *Journal of Wildlife Management*, **64**, 912–923.

Cavalli-Sforza LL, Bodmer WF (1999) *The Genetics of Human Populations*. W.H. Freeman, San Francisco.

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

Chakraborty R, Nei M (1977) Bottleneck effects on the average

heterozygosity and genetic distance with the stepwise mutation model. *Evolution*, **31**, 347–356.

Chakraborty R, Rao CR (1991) Measurement of genetic variation for evolutionary studies. In: *Handbook of Statistics* (eds Rao CR, Chakraborty R), Vol. 8, pp. 271–316. Elsevier Science Publishers, New York.

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

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

Felsenstein J (1985) Phylogenies from gene frequencies: a statistical problem. *Syst Zoology*, **34**, 300–311.

Foulley J, Hill WG (1999) On the precision of estimation of genetic distance. *Genetics, Selection, and Evolution*, **31**, 457–464.

Goldstein DB, Pollock DD (1997) Launching microsatellites: a review of mutation processes and methods of phylogenetic inference. *Journal of Heredity*, **88**, 335–342.

Goldstein DB, Schlotterer C, eds (1999) *Microsatellites*. Oxford University Press, Oxford, UK.

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 of Evolutionary Biology*, **7**, 1–44.

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

Li W, Nei, M (1975) Drift variances of heterozygosity and genetic distance in transient states. *Genetics Research Camb*, **25**, 229–248.

Nei M (1972) Genetic distance between populations. *American 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, Roychoudhury AK (1974) Sampling variances of heterozygosity and genetic distance. *Genetics*, **76**, 379–390.

Nei M, Tajima F, Tateno Y (1983) Accuracy of estimated phylogenetic trees from molecular data. *Journal of Molecular Evolution*, **19**, 153–170.

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

Waples RS (1991) Pacific salmon, *Oncorhynchsus* spp. & the definition of 'species' under the Endangered Species Act. *Mar Fish Review*, **53**, 11–22.

Weir BS (1996) *Genetic Data Analysis II*. Sinauer, Sunderland, MA.

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

Steven Kalinowski is a population geneticist in the Conservation Biology Division of the National Marine Fishery Service's Northwest Fisheries Science Center (*http://www.nwfsc.noaa.gov/*). His work applies genetic data and principles to variety of conservation problems ranging from selecting mating pairs within captive breeding programs to defining evolutionary significant units within species.

Appendix I

History of the chord distance

Cavalli-Sforza & Edwards (1967) represented allele frequencies in a population as a point on a multidimensional hypersphere, and defined the chord distance as the length of a chord joining points corresponding to two populations. For one locus, the chord distance was calculated

$$D_{c(j)} = k\sqrt{1 - \sum_{i=1}^{N_{\text{alleles}}} \sqrt{x_i y_i}} \tag{A1}$$

where $k = 2\sqrt{2}/\pi$, x_i represents the frequency of the i^{th} allele in one population, and y_i the frequency of the allele in the other population. The genetic distance over many loci was originally (1967) calculated using the pythagorean theorem in multiple dimensions

$$D_c = \sqrt{\sum_{j=1}^{N_{\text{loci}}} D_j^2}. \tag{A2}$$

Alternatively, the multilocus chord distance has been calculated as the arithmetic mean of the single locus chord distances (A1) (e.g. Takezaki & Nei 1996)

$$D_c = \frac{\sum_{j=1}^{N_{\text{loci}}} D_{c(j)}}{N_{\text{loci}}}. \tag{A3}$$

The square of the original multilocus chord distance (A2), often represented by f_θ to emphasize its relationship to the kinship coefficient, has also been used (see Cavalli-Sforza & Bodmer 1999) as a genetic distance

$$f_\theta = \frac{4\sum_{j=1}^{N_{\text{loci}}} \sum_{i=1}^{N_{\text{alleles}}} \sqrt{x_i y_i}}{\sum_{j=1}^{N_{\text{loci}}} N_{\text{alleles}}\ 1}. \tag{A4}$$

This genetic distance is related to F_{ST} when there are two alleles at a locus, and has the advantage of increasing fairly linearly with time following isolation of populations. One disadvantage of this genetic distance is that it is expected to decrease with increasing sample size because larger samples are expected to contain more alleles. Nei *et al*. (1983) therefore, defined the D_A distance to alleviate this problem

$$D_A = \frac{\sum_{j=1}^{N_{\text{loci}}} 1 \sum_{i=1}^{N_{\text{alleles}}} \sqrt{x_i y_i}}{N_{\text{loci}}}. \tag{A5}$$

This version of the chord distance (A5) weights the contribution of each locus equally, regardless of the number of alleles present at that locus. This presumably results in a less precise genetic distance.

Unfortunately, the effects of evolutionary parameters upon the four chord distances described here have not been clearly described, especially for loci with a high mutation rate (but see Nei *et al*. 1983; Felsenstein 1985; Cavalli-Sforza & Bodmer 1999). Preliminary work (Kalinowski unpublished) suggests that the original chord distance (A2) and its arithmetic mean formulation (A3) are much less linear than the squared chord distance (A4) and the D_A distance (A5).

Appendix II

Variances associated with sampling loci and individuals

Unlinked loci respond to genetic drift stochastically and independently. One consequence of this is that the genetic distance between populations varies across loci (Fig. 8). The parametric genetic distance between the two populations is a function of the genetic distances at each locus. Therefore, sampling many loci is necessary to obtain good estimates of genetic distance. The variance of the distribution of genetic distances across loci is called the interlocus variance. This variance increases as populations are isolated for longer periods of time (Fig. 8). Interlocus variance, $V_{\text{loci}(m)}$, is one of two contributors to the sampling variance, $V_{\text{sampling}(m,n)}$, of estimates of genetic distances, where m refers to the number of loci sampled and n refers to the number of individuals sampled at each locus. A second source of sampling variance is intralocus variance, $V_{\text{inds}(m,n)}$, which is the sampling variance attributable to sampling a limited number of individuals at the loci being examined. Each of these three variances are related by

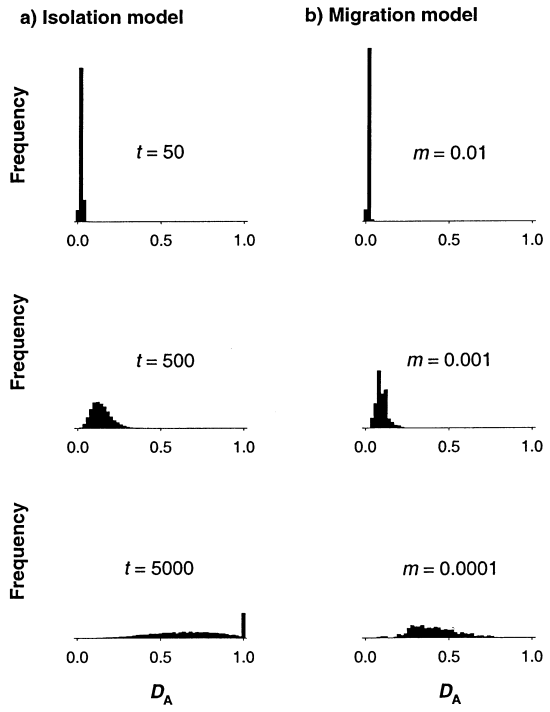

Fig. 8 The distribution of D_A across 10 000 loci for two populations of 5000 individuals isolated for t = 50, 500, or 5000 generations (a) or connected by a migration rate of m = 0.01, 0.001, or 0.0001 (b). The mutation rate is 10^{-4} (iam) in each case. Histograms for D_S, θ are similar (unpublished).

$$V_{\text{sampling}(m,n)} = V_{\text{loci}(m)} + V_{\text{inds}(m,n)}. \quad \text{(A1)}$$

Interlocus sampling variance is reduced by sampling large numbers of loci; intralocus variance is reduced by sampling large number of individuals. The most significant aspect of A1 is that the sampling variance of estimates of genetic distance is always at least as large as the interlocus sampling variance. The consequence of this is that increasing the sample size (i.e. number of individuals sampled) has a diminishing impact upon the sampling variance. If the intralocus variance is small compared to the interlocus variance, then sampling more individuals will not be an effective way to reduce the sampling variance. In this case, sampling more loci will be the only effective way to obtain better estimates of genetic distances. Efficient study design therefore requires knowing the relative magnitudes of the inter- and intralocus variances. Nei and coworkers have examined relationships between inter- and intralocus variances of the genetic distances of Nei (Nei 1972; Nei & Roychoudhury 1974; Li and Nei 1975; Nei 1978).