What is a population?

An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity

Robin S. Waples\textsuperscript{1} and Oscar Gaggiotti\textsuperscript{2}

\textsuperscript{1}Northwest Fisheries Science Center
2725 Montlake Blvd. East
Seattle, WA, 98112 USA

\textsuperscript{2}Laboratoire d’Ecologie Alpine (LECA)
Génomique des Populations et Biodiversité
Université Joseph Fourier, Grenoble, France

Corresponding author:
Robin Waples (robin.waples@noaa.gov)
(206) 860-3254 (voice)
(206) 860-3335 (FAX)

Key words: gene flow, migration, genetic differentiation, demographic independence, ecological paradigm, evolutionary paradigm, \textit{F}-statistics

Running title: What is a population?

\textit{An invited review (in press, Molecular Ecology)}

September 2005
Abstract

We review commonly used population definitions under both the ecological paradigm (which emphasizes demographic cohesion) and the evolutionary paradigm (which emphasizes reproductive cohesion) and find that none are truly operational. We suggest several quantitative criteria that might be used to determine when groups of individuals are different enough to be considered “populations.” Units for these criteria are migration rate \((m)\) for the ecological paradigm and migrants per generation \((Nm)\) for the evolutionary paradigm. These criteria are then evaluated by applying analytical methods to simulated genetic data for a finite island model. Under the standard parameter set that includes \(L = 20\) High mutation (microsatellite-like) loci and samples of \(S = 50\) individuals from each of \(n = 4\) subpopulations, power to detect departures from panmictia was very high (~100%; \(P < 0.001\)) even with high gene flow \((Nm=25)\). A new method, comparing the number of correct population assignments with the random expectation, performed as well as a multilocus contingency test and warrants further consideration. Use of Low mutation (allozyme-like) markers reduced power more than halving \(S\) or \(L\). Under the standard parameter set, power to detect restricted gene flow below a certain level \(X (H_0: Nm < X)\) can also be high, provided that true \(Nm \leq 0.5X\). Developing the appropriate test criterion, however, requires assumptions about several key parameters that are difficult to estimate in most natural populations. Methods that cluster individuals without using \textit{a priori} sampling information detected the true number of populations only under conditions of moderate or low gene flow \((Nm \leq 5)\), and power dropped sharply with smaller samples of loci and individuals. A simple algorithm based on a multilocus contingency test of allele frequencies in pairs of samples has high power to detect the true number of populations even with \(Nm = 25\) but requires more rigorous statistical evaluation. The ecological paradigm remains challenging for evaluations using genetic markers, because the transition from demographic dependence to independence occurs in a region of high migration where genetic methods have relatively little power. Some recent theoretical developments and continued advances in computational power provide hope that this situation may change in the future.
Introduction

A centerpiece of the modern evolutionary synthesis has been development of a rich body of population genetic theory. Early work by Wright, Fisher, and others has been expanded and applied to a vast range of species and biological questions. A recurrent theme of this body of work is the study of genetic structure of species in nature and elucidation of patterns of genetic and demographic connectivity among different groups of individuals, or ‘populations.’ The concept of a ‘population’ thus is central to the fields of ecology, evolutionary biology, and conservation biology, and numerous definitions can be found in the literature (Table 1).

Given the central importance of the population concept, it might be expected that one could take a commonly used population definition and apply it directly to species in the wild to determine how many populations exist and characterize the relationships among them. Furthermore, one might expect that the definition would be objective and quantitative enough that independent researchers could apply it to a common problem and achieve the same results. In fact, however, few of the commonly used definitions of ‘population’ are operational in this sense; instead, they typically rely on qualitative descriptions such as “a group of organisms of the same species occupying a particular space at a particular time” (Krebs 1994; Table 1). It is easy to see that, confronted with a common body of information, different researchers might come to different conclusions about the number of populations and their interrelationships.

Although the difficulties in defining what a population represents have been widely recognized for some time, this problem has, curiously, remained largely unexplored in the literature. Several recent developments indicate that more concerted effort on this issue would be timely. First, availability of numerous, highly polymorphic DNA markers has spurred an explosive interest in genetic studies of natural populations. These studies have considerable power to detect population structure and routinely estimate population parameters without (generally) attempting to define what a population is. Second, new statistical methods, which allow one to identify the number of ‘populations’ in a group of samples and/or assign individuals to population of origin (Paetkau et al. 1995; Rannala and Mountain 1997; Pritchard et al. 2000; Corander et al. 2003), are being widely and energetically applied. In the absence of a common understanding of what a population represents, it can be difficult to evaluate or compare results of such analyses. Third, recent theoretical and empirical studies (Beerli 2004; Slatkin 2005) have reemphasized the point that interactions with unsampled (“ghost”) populations can affect estimates of key parameters (migration rate, population size, genetic diversity) for populations of interest. Evaluating the nature and magnitude of potential biases caused by this phenomenon implies an operational definition of ‘population.’ Finally, genetic data are increasingly being used to inform conservation and management (Moritz 1994; Waples 1995; Crandall et al. 2000; Allendorf et al. 2004). For practical as well as biological reasons, ‘populations’ are natural focal units for conservation and management (McElhany et al. 2000; Beissinger and McCullough 2002), and identification of population boundaries can have far-reaching management (and legal) implications.

To make progress toward resolving these issues, a number of key questions must be addressed. For example, “What is a population (conceptually)?” “Does the variety of population definitions in the literature represent inevitable variations on a common theme, or does it reflect
a fundamental divergence of views regarding what a population is?” “What specific analyses or
tests can be applied to determine whether a unit of interest represents a population?” “How do
these analyses/tests perform with real data, and how does performance depend on the choice of
population definition and criteria to evaluate them?”

To address these questions, a conceptual framework is needed to frame the problem.
Second, it is necessary to define quantitative criteria that can make the conceptual definitions
operational. Third, because it is often difficult to evaluate the criteria directly, metrics must be
developed that can be measured or computed for species in the wild. These metrics can be used
to determine whether population criteria have been met. Finally, analysis of realistic sample
datasets is important to make the examples concrete and evaluate performance of various
population definition, criteria, and metrics.

Collectively, this represents an ambitious research program—much more than can be
accomplished in a single paper. Our objectives here are more limited. First, we briefly review
published definitions of biological “populations” and identify some common themes. Second,
we suggest quantitative criteria and metrics that might be used to make some generic population
definitions operational. Finally, we empirically evaluate performance of a number of genetic
methods for identifying the number of “populations” and their degree of connectivity. Because
the potential parameter space to consider is so large, we have chosen to focus on a relatively
simple model of population structure and assess sensitivity of results to factors of specific
interest to researchers involved in the study of natural populations: type of genetic markers,
numbers of individuals and gene loci sampled, number of populations, and population size.

**Conceptual Framework**

*Population definitions*

Table 1 is certainly not an exhaustive list of population definitions but it is intended to be
representative. As a first cut we can distinguish statistical vs biological definitions. The former
refer to an aggregate of things (which may or may not represent individuals) about which one
wants to draw inferences by sampling. Biological definitions, in contrast, refer exclusively to
collections of individuals that share some biological attributes (but see Pielou 1974 for a largely
statistical definition of a biological population). This paper will be concerned with biological
definitions of ‘population.’

Although a wide range of biological definitions can be found in the literature, some
patterns are apparent. First, all imply a cohesive process that unites individuals within a
population. Second, two major types of biological definition can be identified (Andrewartha and
Birch 1984; Crawford 1984): those reflecting an ecological paradigm and those reflecting an
evolutionary paradigm. Within each paradigm, various flavors of definition can be found, but all
share strong commonalities. In the ecological paradigm, the cohesive forces are largely
demographic, and emphasis is on co-occurrence in space and time so that individuals have an

*1 Numerous variations on population terminology and definitions (e.g., ‘deme,’
‘subpopulation,’ ‘stock’; Table 1) have also appeared in the literature. We will not attempt to
address these terms here, except to note that they could be evaluated using the same general
framework adopted here for ‘population.’
opportunity to interact demographically (competition, social and behavioral interactions, etc.).

In the evolutionary paradigm the cohesive forces are primarily genetic, and emphasis is on
reproductive interactions between individuals. We will consider these two population paradigms
separately, and we adopt a general working definition of “population” for each paradigm as
follows:

Ecological paradigm: A group of individuals of the same species that co-occur in space and
time and have an opportunity to interact with each other.

Evolutionary paradigm: A group of individuals of the same species living in close enough
proximity that any member of the group can potentially mate with any other member.

A simple metapopulation model

We use a simple model to make this problem concrete and allow quantitative analysis.

Consider a metapopulation comprised of \( n \) subunits (subpopulations; \( n \geq 2 \)) that might or might
not represent ‘populations’. Within each subpopulation mating is random, and the
subpopulations are linked (perhaps) by migration. Two extreme scenarios can be identified
(Figure 1). In the first (Figure 1A), the subpopulations are completely isolated (no direct genetic
or demographic linkages) and do not really behave as a metapopulation at all, except perhaps on
very long time scales. In this scenario, therefore, the subpopulations would be considered
separate populations under both paradigms. At the other extreme (Figure 1D), mating is random
within the entire metapopulation; in this scenario, therefore, the metapopulation is panmictic and
the subpopulations are arbitrary. In a metapopulation with \( n \) subpopulations and total size
\( N_T = \sum_{i=1}^{n} N_i \), panmixia occurs when, for each subpopulation, the proportion of migrants is given
by \( m_i = (N_T - N_i) / N_T \)--that is, when the probability of not migrating from the natal
subpopulation \((1-m_i)\) is just the ratio of the size of the natal subpopulation to the metapopulation
size \((N_i/N_T)\). If all subpopulations are the same size, then panmixia occurs when all \( m_i = (n-1)/n \).

Most real-world situations are intermediate to these two extremes (Figure 1B and 1C).
This raises two fundamental questions with respect to population identification. First, given that
the magnitude of departure from panmixia occurs along a continuum (Figure 1, bottom), how
does one define a point along that continuum at which subunits are differentiated enough to be
considered “populations”? With the exception perhaps of McElhany et al. (2000), none of the
definitions in Table 1 is quantitative enough to serve as an unambiguous guide for answering this
question. It will therefore be necessary to consider alternative criteria to make the working
definitions for the two paradigms operational. Second, assuming one has defined a point along
the continuum that corresponds to the concept ‘population,’ how can one in practice determine
whether units of interest are populations? This is a quantitative question that requires developing
population metrics that can be evaluated for power and sensitivity.

Population Criteria

Evolutionary paradigm. Reproductive cohesiveness is determined by levels of gene flow.
As shown by Wright (1931), the evolutionary consequences of gene flow scale with the absolute
number of effective migrants, \( N_e m \), so population criteria under the evolutionary paradigm
should be couched in terms of \( N_e m \). What values of \( N_e m \) might correspond to separate
populations? First, one might consider that separate populations exist when any departure from
Panmixia is found. Assuming an island model in which all migration rates are the same (all \( m_i = m \)), panmixia occurs when \( m = (n-1)/n \), which implies that \( N_e m = N_e(n-1)/n \). That is, in a panmictic metapopulation the number of immigrants per generation into each subpopulation is \( N_e(n-1)/n \). This suggests one possible population criterion:

**Criterion EV1:** \( N_e m < N_e(n-1)/n \).

Another possible criterion depends on the relative importance of migration and drift in determining subpopulation allele frequencies. If \( m \ll 1/N_e \), then the random (dispersive) process of drift dominates and population allele frequencies tend to behave independently. If \( m \gg 1/N_e \), the deterministic (cohesive) force of gene flow dominates, limiting the amount of divergence among subpopulations. A transition between these two regimes occurs at approximately \( m = 1/N_e \), or \( N_e m = 1 \). Therefore, another possible population criterion is:

**Criterion EV2:** \( N_e m < 1 \).

\( N_e m = 1 \) (one migrant per generation) is commonly used as a guideline for management of endangered species (e.g., Mills and Allendorf 1996; Wang 2004). However, EV2 may be too stringent as a population criterion, because substantial departures from random mating (and substantial differences in subpopulation allele frequency) can occur when \( N_e m > 1 \). Choice of any particular value in the range \( 1 < N_e m < N_e(n-1)/n \) is somewhat arbitrary. To capture the range commonly encountered in studies of species in nature, we explore two additional criteria:

**Criterion EV3:** \( N_e m < 5 \)

**Criterion EV4:** \( N_e m < 25 \).

Using the well-known approximation \( F_{ST} \approx 1/(1+4N_e m) \), \( N_e m < 5 \) implies \( F_{ST} > 0.05 \). Wright (1978) indicated that genetic differentiation is “by no means negligible” if \( F_{ST} \) is as small as 0.05. If \( N_e m \) is as large as 25, \( F_{ST} \) will be \( \sim 0.01 \), a small value that nevertheless can be associated with statistically significant evidence for departures from panmixia.

**Ecological paradigm.** Demographic cohesiveness scales with the fraction of the subpopulation that immigrates from other subpopulations (\( m \)). One could test whether \( m \) is less than expected under panmixia (the analogue to Criterion EV1 is \( m < (n-1)/n \)), but such a test has limited relevance for most ecological considerations. A more relevant question is, How small must \( m \) be before the subpopulations are demographically independent? Although this question would appear to be fundamental to understanding metapopulation processes, it has apparently received little formal study. The limited available information (Hastings 1993) suggests that transition to demographic independence occurs when \( m \) falls below about 10%. This suggests a possible criterion:

**Criterion EC1:** \( m < 0.1 \).

As discussed in Methods, we considered several different metrics to test whether these population criteria are met and evaluated their performance using simulated data.

**Methods**

**Simulated data**

Genotypic data were generated by EASYPOP (Balloux 2001). We considered a finite island model with \( n \) subpopulations, each of constant size \( N \) and equal sex ratio. Each generation, random mating was simulated to produce a diploid genotype for \( L \) independent gene
loci for each individual, which then had probability \( m \) of migrating to another subpopulation. Under this Wright-Fisher process, \( N_e \approx N \) in every subpopulation. In the following, therefore, we will use the term \( Nm \) to represent the effective number of migrants per generation \((N_e m)\). Within a parameter set, all loci had the same mutation dynamics, which occurred according to the \( K \)-allele model (KAM; each mutation equally likely to occur at any of \( K \) possible sites). Two combinations of mutation rate \((\mu)\) and number of possible allelic states were considered, one representative of highly polymorphic markers like microsatellites (Estoup and Angers 1998; \( \mu = 5 \times 10^{-4}; 10 \) allelic states), the other representative of low-mutation rate markers like allozymes or single-nucleotide polymorphisms (SNPs) (Zhang and Hewitt 2003; Morin et al. 2004; \( \mu = 5 \times 10^{-7}, 4 \) allelic states). In what follows, we will refer to these two mutation patterns as “High” and “Low,” respectively. Simulations were initiated with maximal genetic diversity (genotypes in initial generation randomly drawn from all possible allelic states). Although the magnitude of population differentiation reaches equilibrium rapidly under the conditions considered here (Crow and Aoki 1984), we ran each replicate for 5000 generations before collecting data to attain an approximate mutation-drift equilibrium. In the final generation of each replicate, samples of \( S \) individuals were taken from each subpopulation for genetic analysis. Default values for key parameters (the “standard model”) were: \( N = 500 \), \( n = 4 \), \( S = 50 \), \( L = 20 \), High mutation, and \( m \) was chosen to yield \( Nm \) values ranging from 0.1 individual/generation to panmixia. Except as noted, we analyzed 100 replicates for each parameter set (Table 2). Each parameter set was given a two-part name, with the second part indicating the number of migrants per generation \((Nm)\) and the first part indicating changes from the standard parameter set (Hi = standard set with High mutation markers; Lo = standard set with Low mutation markers; \( 25S = \) sample size of 25; \( 10L = 10 \) loci; \( 2n, 8n = 2 \) or 8 subpopulations; \( 200N, 100N, 50N = \) subpopulation size different than 500; \( C = \) combination low power with Low mutation markers, \( L = 10 \), and \( S = 25 \).

Testing panmixia

Contingency tests. Contingency tests of allele frequency heterogeneity followed the method of Raymond and Rousset (1995), which uses Markov Chain Monte Carlo (MCMC) methods to provide an unbiased estimate of the exact probability for each single-locus comparison. Calculations were performed using a version of the program RXC (available at http://www.marksgeneticsoftware.net/Miller program) that was modified to a) allow batch processing of multiple datasets, and b) compute a multilocus \( P \) value for each comparison using Fisher’s method for combining probabilities across loci. For each randomization test, we ran 10 batches of 10,000 replicates each, with 1000 dememorization steps. To minimize opportunities for a single locus to dominate the overall test (Lugon-Moulin et al. 1999), we constrained the single locus \( P \) values to be no smaller than 0.0001.

Assignment tests. Assignment tests used the Rannala and Mountain (1997) method as implemented in GENECLASS2 (Piry et al. 2004). An individual was considered correctly assigned if assignment was to the population in which it was sampled. First-generation migrants might be incorrectly assigned by this criterion. With \( N = 500 \) individuals per subpopulation, \( Nm = 1, 5, \) and 25 migrants per generation represented 0.2%, 1%, and 5% of each subpopulation, respectively. Therefore, the maximum expected percentage of correct assignments was 99.8%, 99%, and 95%, respectively, for the three levels of migration. The observed percentage of correct assignments was averaged over all subpopulations within a replicate and then across all replicates within a parameter set. For each replicate, the number of correct assignments was
compared with that expected under random assignment as follows. If there are $n$ potential sources represented by samples of equal size, the probability of correctly assigning at random any given individual is $p = 1/n$. If the total number of individuals to be assigned is $N_A = nS$, then the expected number of random, correct assignments is $nS/n = S$. The probability of a specific number $X$ of correct assignments at random is given by the binomial distribution:

$$Pr(#\text{correct} = X | N_A, p) = \frac{N_A!}{X!(N_A - X)!} p^X (1 - p)^{N_A - X}. \quad (1)$$

To evaluate whether the observed number of correct assignments was significantly higher than the random expectation, we used the cumulative binomial distribution to identify critical values for the number of correct assignments. Results for parameter sets considered here are shown in Table 3. For example, in the standard parameter set with $n = 4$, $S = 50$ (hence $N_A = 200$) 60 or more correct assignments is significant at the $P<0.05$ level, 65 is significant at the 0.01 level, and 70 correct assignments are needed to demonstrate performance better than random at $P<0.001$.

**F statistics.** The most commonly used measure of genetic differentiation among populations is Weir and Cockerham’s (1984) $\theta$, an analogue to $F_{ST}$. To obtain expected values of $\theta$ for different combinations of parameters $m$, $u$, $n$, and $N = N_e$, we used the formula of Cockerham and Weir (1987, 1993), which assumes that $u$ and $m$ are small:

$$E(\theta) \approx \frac{1}{1 + 4Nu + 4NmN/(n-1)}.$$

The relationship between $\theta$ and $G_{ST}$ (the multilocus version of $F_{ST}$) is $\theta = nG_{ST}/(G_{ST}+n-1)$ (Cockerham and Weir 1987, 1993). If one makes this substitution for $\theta$ in Equation 2 and assumes that mutation is low enough to be ignored, the result is

$$G_{ST} \approx \frac{1}{1 + 4Nm(n/(n-1))^2},$$

as obtained by Crow and Aoki (1984). If one further assumes that the number of subpopulations is large enough that the term $n/(n-1)$ can be ignored, one obtains Wright’s familiar formula,

$$G_{ST} \approx \frac{1}{1 + 4Nm}.$$

We used FSTAT (version 2.9.3.2; Goudet 1995) to calculate Weir and Cockerham’s estimator $\hat{\theta}$, confidence intervals (CIs) for $\hat{\theta}$ by bootstrapping over loci, and average gene diversities ($H_s = \text{expected heterozygosity averaged across subpopulations};$ Nei 1987).

Expected values of $\theta$ for three different $Nm$ values (1, 5, and 25, corresponding to critical values for Criteria EV2-4) were computed for each parameter set using Equation 2, and these were used as critical values to test hypotheses about gene flow. For example, assume we want to test the hypothesis that gene flow is less than 25 migrants per generation (Criterion EV4; $H_0$: $Nm<25$), given the following parameter values: $n = 4$; $N = 500$; $\mu = 0.0005$. With $N = 500$, $Nm = 25$ implies $m = 0.05$, and inserting these values for $n$, $N$, $\mu$, and $m$ in Equation 2 yields $E(\theta) = 0.0074$. If the lower CI of an observed $\hat{\theta}$ is greater than the critical value 0.0074, it can be concluded that gene flow is unlikely to be as high as $Nm = 25$. 

8
Estimating the number of populations

In these evaluations, the fraction of replicates for which the estimated number of populations \( \hat{k} \) was equal to the true \( n \) was used as a performance measure.

Putative populations defined a priori. We compared the performance of two programs that assume each sample is drawn from only one population, but that some populations might have been sampled more than once. In these tests, therefore, the number of samples represents an upper limit for \( \hat{k} \).

We used RXC as described above to identify replicates in which homogeneity among all the samples could not be rejected at \( P<0.01 \); these replicates were considered to include just one population \( \hat{k} = 1 \). For replicates showing overall heterogeneity, the number of different populations represented by the \( n \) samples was calculated in the following way. First, RXC was used to test whether allele frequencies in each of the \( J = n(n-1)/2 \) pairwise comparisons differed at the \( P<0.01 \) level. Next, a link was drawn between all pairs of samples not differing significantly (see Figure 2). A group of samples was considered to come from the same population if every pair within the group could be connected through a chain of non-significant tests. In the example in Figure 2, \( n = 8 \) samples are determined to represent 3 populations; population A is comprised of a single sample that differs significantly from all others, whereas populations B and C include 4 and 3 linked samples, respectively.

Because the pairwise RXC method involves multiple tests within each replicate (the number of pairwise comparisons is \( J = 1, 6, \) and 28 for \( n = 2, 4, \) and 8, respectively), a certain fraction is expected to be significant just by chance. Quantitative adjustment for multiple testing is problematical because the different pairwise tests are not independent. Nevertheless, some insight into the magnitude of the potential problem can be gained by treating the comparisons as if they were independent. In that case, under panmixia the probability that none of the pairwise tests within a replicate is significant is \( (1-\alpha)^J \); for \( \alpha = 0.01 \) this probability is over 94\% for \( n = 4 \) and over 75\% for \( n = 8 \). Assuming independence, the chances that all pairwise tests will be significant by chance \( (\alpha^J) \) is very remote for \( n > 2 \). We therefore expect that under conditions considered here, multiple testing issues will not strongly affect results of the RXC method to estimate the number of populations. In Results we present empirical data from the simulations that bear on this issue.

We also evaluated the “cluster groups of individuals” option of BAPS (version 3.1; Corander et al. 2003; available from http://www.rni.helsinki.fi/~jic/bapspage.html), which uses a Bayesian approach to determine which combination of predetermined samples is best supported by the data. \( \hat{k} \) was taken to be the partition with the highest posterior probability. The program uses importance sampling to approximate posterior probabilities for large datasets, but for \( n \leq 8 \) (as considered in this study), BAPS performs an exact Bayesian analysis by enumerative calculation to arrive at \( \hat{k} \).

Putative populations not defined a priori. The estimation procedure for STRUCTURE 2.0 (Pritchard et al 2000) consists in running the program for different trial values of the number of populations, \( k \), and then comparing the estimated log probability of the data under each \( k \).
ln[Pr(X|k)].  \(^\hat{k}\) was taken to be the value with the highest Pr(X|k).  A pilot study indicated that runs with a burn in of 30,000 and a total length of 100,000 provided consistent estimates of Pr(X|k) when genetic differentiation was strong to moderate (Nm = 1–5). However, we were unable to obtain convergence when genetic divergence was low (Nm = 25), even for runs of up to 4 million iterations. We chose the admixture model and the option for correlated allele frequencies, both appropriate for the migration model we used. For each parameter set we analyzed 10 replicate data sets and recorded the proportion of correct assignments and Pr(X|k).

Evanno et al. (2005) suggested that an ad hoc measure, \(\Delta k\), the second order rate of change of ln[Pr(X|k)] with respect to k, provides a more reliable estimator. This measure was calculated by carrying out many trial runs of STRUCTURE (e.g., 20) for each putative k value in each replicate data set and then applying the following equation:

\[
\Delta k = \frac{\text{mean}(|\text{Pr}(X|k+1) - 2\text{Pr}(X|k) + \text{Pr}(X|k-1)|)}{\text{sd}[\text{Pr}(X|k)]},
\]

where mean represents the mean and sd represents the standard deviation across trials. Due to computational constraints we adopted this procedure only for a limited number of scenarios and used only 5 trial runs of STRUCTURE for each replicate dataset.

Results

Levels of genetic variability

In simulations using High mutation, all or nearly all loci were polymorphic (two or more alleles in at least one sample; Table 2). Occasional exceptions occurred with \(N \leq 200\) or \(n = 2\), in which case the overall metapopulation size was relatively small and some loci drifted to fixation. Under the “standard” model (\(n = 4, N = 500\), High mutation), average subpopulation gene diversities were \(H_s \sim 0.7\) (Table 2), comparable to values commonly reported in studies of natural populations using microsatellite markers. Levels of variability were only about half as high in simulations using Low mutation (\(H_s \sim 0.35\)), and only about 2/3 of the loci were polymorphic (Table 2). Still, the levels of variability were at least as high as those reported in most allozyme studies of natural populations (e.g., Figure 10 in Hartl and Clark 1988).

Type I error rates

Before analyzing population subdivision, we evaluated Type I error rates under conditions in which the entire metapopulation was panmictic. We used standard parameter sets Hi-P (High mutation) and Lo-P (Low mutation) and evaluated 1000 (rather than 100) replicate datasets. The multilocus contingency test produced almost exactly the expected number of significant tests at each significance level (Appendix Table 1): at the \(P<0.05\) level, 49 tests were significant for High mutation and 50 for Low mutation (50 expected); at the \(P<0.01\) level, 9 (High) and 10 (Low) were significant (10 expected); at the \(P<0.001\) level, 1 (High) and 0 (Low) were significant (1 expected). We also found general agreement between the observed and expected distribution of multilocus P values over the full range 0-1 (\(P>0.05\) for both High and Low mutation markers; Kolmogorov—Smirnov goodness of fit test). Testing panmixia by comparing observed numbers of correct assignments with the random expectation resulted in slightly elevated Type I error rates under both High and Low mutation for each nominal \(\alpha\) level considered (Appendix Table 1). However, the mean percentage of correct assignments (24.9% for High mutation; 24.6% for Low mutation) was very close to the random expectation (25% with \(n = 4\)).
Bootstrapped CIs for \( \hat{\theta} \) performed somewhat erratically. Under the standard parameter set (High mutation), the lower 95% CI should be larger than zero 2.5% of the time and the lower 99% CI should be larger than zero 0.5% of the time; the observed rates of Type I error (9.3% and 2.3%, respectively; Appendix Table 1) were 3-5 times as high as expected. A similar, although slightly less pronounced, upward bias in the Type I error rate was found with Low mutation markers. In the case of \( \hat{\theta} \), it is also possible to test conformance with null hypothesis expectations for non-zero levels of gene flow, based on comparing observed \( \hat{\theta} \) values with those expected using Equation 2. This allowed evaluation of the CIs for \( \hat{\theta} \) for a variety of parameter sets with true \( Nm = 25, 5, \) or 1. Results (bold cells in Appendix Table 1) varied across parameter sets, with the following general tendencies: the test was slightly conservative (rejecting \( H_0 \) less often than expected) with \( Nm = 25 \) but had approximately the expected Type I error rate for \( Nm = 5 \) or 1; and Type I error rates were slightly elevated for parameter sets using fewer loci and/or smaller samples.

**Evolutionary paradigm**

**Testing departures from panmixia.** As shown in Appendix Table 1, all three methods performed well in detecting departures from panmixia, even for “hard” problems with low levels of genetic differentiation. For example, with the standard parameter set and \( Nm = 25 \) (Hi-25), all three methods detected significant population structure 100% of the time using the most stringent criterion (\( P < 0.001 \) for contingency tests and assignment tests and \( P < 0.01 \) for \( \hat{\theta} \)). As expected, as the problems became even harder (lower mutation rates, fewer loci and populations, smaller sample sizes), performance of all three methods declined somewhat, but performance deteriorated substantially only in the dataset (C-25) that combined all of these factors that reduce power (Appendix Table 1). Over a wide range of “hard” parameter sets, the contingency test and the assignment test methods consistently showed slightly higher power to detect departures from panmixia than did the tests based on CIs for \( \hat{\theta} \) (Figure 3). Of the former two tests, in some cases the contingency test performed slightly better and in other cases the assignment test method had higher power.

**Testing hypotheses about gene flow.** In spite of the somewhat erratic Type I error rate for the method using CIs for \( \hat{\theta} \), agreement between \( \hat{\theta} \) and \( E(\theta) \) was very good for most parameter sets (Appendix Table 1). As expected, given that the approximation in Equation 2 assumes migration and mutation rate are small, proportional deviations from \( E(\theta) \) were slightly larger for large \( m \) values.

Results in Appendix Table 1 also show that under all parameter sets examined, power to detect restricted gene flow (Criteria EV2-4) can be nearly 100 percent, provided that actual \( Nm \) is much lower than the hypothesized level, \( Nm(H) \). For example, under parameter set 10L-5 (true \( Nm = 5 \) and only 10 loci used), in 100% of the replicates the lower 99% CI for \( \hat{\theta} \) was higher than the expected value of \( \theta \) for \( Nm(H) = 25 \) (\( E(\theta) = 0.0349 \) from Equation 2). Thus, if one has data for 10 microsatellite loci in samples of 50 individuals each drawn from populations among which the actual level of gene flow is 5 migrants per generation, one could be very confident in concluding that gene flow must be less than \( Nm = 25 \).
To evaluate in more detail the transition from low to high power to detect restricted gene flow, we conducted additional simulations using the standard model with both High and Low mutation and chose \( m \) to produce realized \( Nm \) values of 20, 15, and 10. In each case we calculated empirical CIs for \( \hat{N}m \) and asked whether the lower CI was higher than \( E(\theta) \) for \( Nm(H) = 25 \) (Criterion EV4). Results (Figure 4) show that with High mutation markers, power to test Criterion EV4 increases rapidly as true \( Nm \) drops below 20 migrants per generation and is \( >90\% \) if \( Nm \) is as low as 10. With Low mutation markers, power remains relatively low unless \( Nm < 10 \). Figure 5 shows a more general result for High mutation markers: the transition from low to high power for a wide range of \( Nm \) values occurs at approximately true \( Nm = 0.5*Nm(H) \); that is, power to detect restricted gene flow is very high if true \( Nm \) is no more than half the hypothesized level, but is low otherwise. For the same ratio of true \( Nm : Nm(H) \), power is slightly higher when \( Nm \) is low. If Low mutation markers are used, power is low unless \( Nm(H) \) is about five times the true \( Nm \) (Figure 4; Appendix Table 1; unpublished data).

As expected, the percentage of correct assignments increases sharply as gene flow becomes more restricted. However, performance of assignment tests also depends heavily on mutation rate and less strongly on \( S, N, n, \) and \( L \) (Figure 6; Appendix Table 1).

**Estimating the number of populations.** The two methods that depend on a priori information about geographic sampling showed dramatically different performance in estimating the true number of gene pools. The pairwise RXC test consistently detected all or nearly all of the populations, except under conditions (C-5) with the lowest cumulative power (Figures 7 and 8). In contrast, BAPS almost always underestimated the true number of populations, often dramatically, except in the case of the most extreme population differentiation (\( Nm = 1 \)).

In Methods we discussed multiple testing issues associated with the pairwise RXC method and concluded that this issue was not likely to strongly affect results of this study. To evaluate this empirically, we considered results for parameter set Hi-P (standard model with 4 samples from a globally panmictic population). Only 9 of 1000 replicates (0.9\%) showed significant heterogeneity at the \( P<0.01 \) level (Appendix Table 1), and in each of those replicates multiple pairwise comparisons had \( P \) values larger than 0.01, leading to \( \hat{k} = 1 \) according to the criteria outlined in Methods and depicted in Figure 2. Therefore, only a single population was detected in each of the 1000 replicates, resulting in an empirical Type I error rate of 0. These results suggest that, at least for relatively small \( n \), the test is conservative and multiple testing issues are not responsible for the observed power of this approach to detect the true number of populations.

**STRUCTURE** proved to be reliable at estimating the true number of populations when gene flow was relatively low (\( Nm \leq 5 \)) and full samples of individuals and highly polymorphic loci were used (Figures 7 and 8). Performance was much worse (\( \hat{k} = true n \) in less than 40\% of replicates) when sample size or the number of loci used was reduced, and **STRUCTURE** did not provide any useful information about the number of populations when gene flow was high (\( Nm = 25 \)) or Low mutation markers were used (Figure 7; Appendix Figure 1).

We did not find the alternative approach to estimating \( k \) proposed by Evanno et al. (2005) to be an improvement over the standard approach (Pritchard et al. 2000) under conditions used here. Both methods performed well when genetic differentiation was strong (\( Nm = 1 \)) and poorly
when differentiation was weak \((Nm = 25)\), but under moderate genetic differentiation \((Nm = 5)\) the standard approach performed better (correct number of populations identified in 90% of replicates vs. 70% for the \(\Delta k\) method; Appendix Figures 1-2 and unpublished data). Given these results and the computational burden imposed by the Evanno et al. procedure (it requires many trial runs of STRUCTURE for each \(k\) value in each replicate), we used the standard procedure for the remainder of the STRUCTURE analyses.

The ability of STRUCTURE to correctly assign individuals to population of origin is lower than that of the classical assignment test, and the proportional difference increases as the problems become harder (higher \(Nm\); fewer loci and individuals; Low mutation: Figure 9).

**Ecological paradigm**

Statistical tests of population differentiation proved to have high power over a wide range of migration rates. Regardless which test was used (contingency test, assignment test, CI for \(\hat{\Theta}\)), power to detect highly significant population structure was 100% or nearly so for migration rates that spanned the range \(m = 0.0002\) to 0.1 (Table 2 and Appendix Table 1.) Even with \(m\) as high as 0.2 (twice as high as Criterion EC1 for demographic independence), under the standard model \(\text{RXC}\) detected significant differentiation at the \(P<0.05\) level in over half the replicates, and over a third of the replicates showed differentiation at the \(P<0.01\) level.

**Discussion**

Our brief review of literature definitions of “population” makes evident a point that should surprise no one: there is no single “correct” answer to the question, “What is a population?” Instead, the answer depends on the context and underlying objectives. Researchers interested primarily in the interplay of different evolutionary forces (selection, migration, drift) will typically favor a population concept couched in terms of reproductive cohesion, whereas those concerned primarily with conservation or management are more likely to be interested in demographic linkages and the consequences of local depletions. Similarly, regardless which population paradigm is adopted, the question, “How different must units be before they can be considered separate populations?” does not have a unique answer; reasonable arguments can be advanced for using any of a variety of points along the continuum of population differentiation as a criterion.

These realities have both desirable and undesirable consequences. The flexible nature of the population concept means that it can be applied to a wide range of scenarios faced by ecologists and evolutionary biologists. On the other hand, this flexibility also can foster ambiguity and confusion among scientists using different population concepts and/or criteria. These difficulties are not unlike those that for many years have surrounded the problem of how to define species (Mayden 1997; Wilson 1999; Wheeler and Meier 2000). The “species problem” involves both conceptual differences and the inherent biological fuzziness of species in nature (Hey et al. 2003), but neither of these factors need represent an insurmountable obstacle to practical application of species concepts.

Although we don’t presume to have a solution to the comparable difficulties associated with the “population problem,” we believe that meaningful dialogue on these issues is more
likely to occur if researchers a) take time to reflect on how their study fits into a conceptual
framework for defining populations; and b) clarify in their publications which population
paradigm they are following and justify choice of specific quantitative criteria for identifying
populations. Toward those ends, we have outlined a basic framework for considering questions
about populations, and we have suggested some possible quantitative criteria for each of the
population paradigms. If this paper generates more awareness and consideration of these issues,
then one of our major objectives will have been accomplished.

A second major objective was to quantitatively evaluate performance of some commonly
used methods for detecting population structure, and results of those analyses are discussed
below.

Levels of variability

With Low mutation markers, a sharp change in patterns of genetic diversity was seen in
the parameter set with the most restricted gene flow (Lo-01; \( Nm = 0.1 \)); in this case, nearly all
loci were polymorphic (\( L_p = 18 \) compared with \( L_p = 12-14 \) for higher \( Nm \); Table 2) but average
subpopulation gene diversity was low (\( H_s = 0.18 \) compared with \( H_s = 0.32-0.36 \) for higher \( Nm \)).
This reflects the observation (Wright 1931) that when \( Nm < 1 \), alleles tend to drift to fixation in
subpopulations, thus lowering \( H_s \). On the other hand, by chance different alleles often become
fixed in different subpopulations, thus “freezing” genetic diversity and maintaining a high level
of polymorphism across the metapopulation as a whole. A similar reduction in \( H_s \) is seen in the
parameter set Hi-01 (Table 2), although with High mutation the effect is more muted because
new alleles are constantly being generated within subpopulations. This phenomenon of
“freezing” diversity is responsible for the conclusion (Wright 1943) that population subdivision
increases overall effective size of the metapopulation. However, this conclusion depends on the
assumption that \( N \) is constant over time, in which case every subpopulation is effectively
immortal (Waples 2002). If subpopulation extinction is allowed, results can be very different.

Testing panmixia

Goudet et al. (1996) considered power of single-locus tests of population genetic
differentiation and found that exact contingency tests and methods based on analogues to \( F_{ST} \): a)
rejected the null hypothesis of no differentiation close to the expected 5% of the time when the
global population was panmictic, and b) had comparable power when sample sizes were equal.
Results presented here extend these conclusions to the case of multiple loci and different \( \alpha \) levels
(\( \alpha = 0.05, 0.01, 0.001 \)). For the multilocus test, we found better agreement with the nominal
Type-I error rate, and slightly higher power, for \( RXC \) than \( \hat{\Theta} \) (Appendix Table 1; Figure 3).
Although we only evaluated balanced sampling, Goudet et al. (1996) found that power decreases
considerably, and more so for \( F_{ST} \) than the contingency test, if sample sizes differ. Fisher’s
method for combining probabilities over independent tests (used here in the multilocus \( RXC \) tests)
can lead to biases in some cases (Goudet 1999; Ryman and Jorde 2001; Whitlock 2005). The \( ad \)
hoc lower limit of \( P \geq 0.0001 \) we placed on single-locus \( P \) values was intended to minimize such
problems, and based on the excellent agreement with nominal Type-I error rates for the \( RXC \) tests
it appears to have been effective for the experimental conditions used here. Nevertheless, those
interested in testing panmixia with multilocus genetic data might want to consider the standard
method of summing chi-square values across loci (Ryman and Jorde 2001), a multilocus
generalization of Goudet et al’s $G$-test implemented by Petit et al. (2001), or the weighted $Z$-method for combining probabilities described by Whitlock (2005).

It therefore seems that a non-parametric approximation to the exact, multilocus contingency test is the most appropriate method for statistical tests of population differentiation. This test can be very powerful even with weak population differentiation. For example, with samples of $L = 20$ microsatellite-like loci and $S = 50$ individuals/population, power to reject panmixia at the $P < 0.001$ level was 100% even with high gene flow ($Nm = 25$) and, consequently, a very small $\hat{\theta}$ (0.006) (Appendix Table 1). This level of data collection is achievable in many contemporary studies of natural populations. Only for parameter set C-25, with reduced samples of individuals and loci and Low mutation markers, was power appreciably diminished. In this study, we have assessed power as a function of the number and type of gene loci, which together are proxies for what is probably a more direct determinant of statistical power – the total number of alleles for which data are available (Kalinowski 2002, 2004; Balding 2003).

Somewhat surprisingly, we found that a very different type of test—based on comparing observed and expected numbers of correctly assigned individuals—performed very similarly to the exact RXC test. Although it was recently suggested (Manel et al. 2005) that a test that takes advantage of multilocus genotypic information might be more powerful than standard tests that focus on gene loci individually, to our knowledge this approach has not been evaluated previously. Our results suggest that this method merits further consideration, particularly because of an indication that it may have higher power than the contingency test under data-poor conditions. One caveat: the values in Table 3 (critical number of correct assignments for nominal $\alpha$ levels) are straightforward to calculate if all samples are of equal size but more complicated when sampling is unbalanced.

Direct comparison of the percentage of correct assignments in our results with those reported by Cornuet et al. (1999) is difficult because the latter study did not consider migration (only different times of isolation) and only evaluated the case of $n = 10$ subpopulations and $N = 1000$. Nevertheless, Cornuet et al. (1999) found that ~100% correct assignments can be obtained using Rannala and Mountain’s (1997) method with $S = 30-50$, $L = 10$ microsatellite loci, and $FST \approx 0.1$ (compare with results for parameter sets Hi-1, 25S-1, and 10L-1, which show the percent correct assignments ranging from 98% to 100% for simulations with $S \geq 25$, $L \geq 10$, and $\hat{\theta} \approx 0.13$; Appendix Table 1).

It should be recognized that the high power to detect small departures from panmixia is something of a two-edged sword: if the test can detect very weak population structure, it can also confuse small artifacts (e.g., non-random sampling, family structure, data errors) with a true signal of population differentiation (Waples 1998). As a consequence, various sources of noise that might otherwise be safely ignored assume a relatively greater importance. This reality argues for careful attention to experimental design, sampling protocols, and data quality control. Furthermore, it emphasizes the importance of understanding the biology of the target species so that potential sampling artifacts can be avoided as much as possible.
Estimating the number of populations

The Bayesian approach for clustering groups of individuals implemented in BAPS proved to be very conservative in identifying population structure; different gene pools could only be detected reliably under very restricted migration (\(Nm = 1; \hat{\theta} > 0.13\)). The reason for this is not clear; possible explanations include: a) the penalty in BAPS for postulating additional populations (and hence estimating additional parameters) is too severe; or b) recent migrants might have obscured differences among populations (J. Corander, pers. comm.). When we used the “cluster individuals” option (in which case the analysis is similar to that performed by STRUCTURE) and \(Nm = 5\), BAPS was more reliable at estimating the true number of populations, with performance comparable to that of STRUCTURE (unpublished data).

In contrast, pairwise, multilocus contingency tests proved to be quite powerful at estimating the number of populations. Across all replicates, 100% of the populations were detected (every pairwise RXC test significant at the \(P<0.01\) level) under the standard parameter set with \(n = 2, 4, \) or 8 populations and \(Nm \leq 5\), even with reduced samples of loci and individuals (Figure 7). With High mutation markers and high gene flow (\(Nm = 25\)) or Low mutation markers and more restricted gene flow (\(Nm = 5\)), all of the pairwise comparisons were significant in at least 70% of the replicates. Results for the panmictic datasets indicate that this result reflects real power to detect population structure rather than an inflated Type I error rate. With respect to the questions of primary interest here, the most important concern regarding multiple testing is not minimizing the familywise error rate (FWER; the probability of even a single false positive test), which is typically accomplished by a Bonferroni correction (e.g., Rice 1989), but rather the false discovery rate (FDR; the fraction of tests in which the null hypothesis is falsely rejected; Benjamini and Hochberg 1995). The FDR recaptures much of the power sacrificed by Bonferroni approaches, especially when a large number of hypotheses are tested (Garcia 2004; Verhoeven et al. 2005), and certain types of positive dependence among the tests can be accommodated (Benjamini and Yekutieli 2001). Even after adjusting for multiple testing, however, to estimate the number of discrete populations requires a set of rules to integrate information from the \(n(n-1)/2\) pairwise comparisons of samples. Figure 2 illustrates one possible ad hoc algorithm, but this topic clearly merits more rigorous evaluation.

When it is not possible to partition individuals into a priori samples (or when the basis for doing so is of uncertain validity), it is necessary to use an approach that clusters individuals without reference to sample information. We chose the most widely used clustering program (STRUCTURE) to represent this class of analyses. The authors (Pritchard et al. 2000; Falush et al. 2003) admit that the procedure to estimate the number of populations is ad hoc and recommend that it be used only as a guide, but these caveats are often ignored. Previous assessments of the performance of STRUCTURE (Evanno et al. 2005) have focused on situations involving strong differentiation. In agreement with those results, we found that STRUCTURE accurately identified the number of populations when \(Nm\) was 5 or lower, mutation was High, and full samples of loci and individuals were used, but performance deteriorated sharply under less ideal conditions (Figure 7). The complete inability of STRUCTURE to correctly estimate the true number of populations using Low mutation markers is somewhat surprising but in agreement with previous observation regarding the factors primarily responsible for statistical power to detect population differentiation. Reduced samples of loci and individuals also affected performance, although not as dramatically as did the type of markers. We note, however, that (assuming \(Nm\) is low enough
to permit adequate resolution), high power can be achieved using a sampling regime \((L = 20 \text{ and } S = 50)\) that is within the range achievable by many molecular ecology laboratories.

The method we found to be most powerful for identifying the number of populations (a simple algorithm based on the multilocus contingency test) is also the least sophisticated. However, caution must be used in comparing this test with approaches that cluster individuals rather than samples, because performance of the former depends on the premise that each sample has been taken randomly from a single population. RXC (or any other method based on comparison of \textit{a priori} samples) cannot detect hidden structure within samples and can produce misleading conclusions if any of the samples include individuals from more than one biological unit.

None of the methods adequately estimated the true number of populations with Low mutation markers and small samples of loci and individuals. This result should be a caution to those wanting to draw inferences about the number of gene pools based on limited data.

Comparison of our results with those of Evanno et al. (2005) highlights the importance of including datasets with weak genetic differentiation in sensitivity analyses. Evanno et al. found that \(\Delta k\) performed better than the original approach proposed by Pritchard et al. (2000) for estimating the true number of populations. However, Evanno et al. only considered scenarios with strong genetic differentiation \((F_{ST} = 0.15-0.4)\)—much higher than the range considered in our analyses of STRUCTURE \((\hat{\theta} = 0.005-0.136)\). Levels of differentiation we considered are within the range of values observed for the majority of natural populations that have been studied (e.g. Bohonak 1999; Fig. 1). Therefore, results from simulation studies that only consider strong genetic differentiation can lead to conclusions about performance that are overly optimistic for many realistic applications. However, because we only considered a simple island model of migration (Evanno et al. considered hierarchically structured populations) and used relatively few trials of STRUCTURE for each \(k\) value, our results comparing the two methods should be regarded as preliminary. Indeed, the \(\Delta k\) approach may work best with population structures other than the island model (J. Goudet pers. com.)

An important point to keep in mind is that a large variance in \(\ln[\Pr(X|k)]\) across different trial runs indicates that the MCMC chain has not converged. We found a large variance in \(\ln[\Pr(X|k)]\) among trials to be common in datasets with weak genetic differentiation. This result argues for considerable caution when interpreting the results of clustering programs such as STRUCTURE for species whose biology suggests high dispersal abilities. Since convergence of the chain depends on characteristics of the dataset being analysed, the best practice is to compare results for replicate runs. If results are not consistent, the length of the chain should be increased; if all efforts fail to result in convergence, this should be reported with the results.

\textit{Testing levels of gene flow}

When the operational population concept requires more than simply testing for panmixia, methods based on CIs for \(\hat{\theta}\) or related indices can be used to test specific hypotheses about restrictions to gene flow. As shown in Figure 5, these tests can also have high power provided that the true level of gene flow is no more than about half of the critical level (the difference must be larger if Low mutation markers or restricted samples of individuals or loci are used).
These tests require that one postulate a value for $E(\theta)$ corresponding to the hypothesized level of gene flow one wants to evaluate. Because $E(\theta)$ depends on mutation rate, particularly when migration is low (Balloux and Lugon-Moulin 2002), these tests can in theory alleviate some of the problems associated with interpretation of highly variable markers pointed out by Hedrick (1999). However, several important caveats need to be mentioned.

First, Equation 2 provides an approximation for $E(\theta)$ based on a simple migration model under the assumption that $m$ and $u$ are “small.” Some features of the island model are relatively robust to violation of underlying assumptions (Rousset 2003), but it is widely recognized that in some cases $F_{ST}$ and analogues can provide misleading information about migration and gene flow (Waples 1998; Whitlock and McCauley 1999), particularly when migration is unbalanced. Furthermore, $E(\theta)$ depends on several key parameters ($N, u, n$) whose true values are generally unknown. In our model, the number of subpopulations sampled was the same as the true number ($n$), but often this will not be the case. Unsampled “ghost” populations can affect gene flow estimates among the sampled populations in complex ways (Beerli 2004; Slatkin 2005). Collectively, these factors mean that in practice it will be difficult to obtain a reliable $E(\theta)$ for testing a particular level of gene flow.

Second, Equation 2 assumes an equilibrium between drift, mutation, and migration. Although $F_{ST}$ and $\theta$ approach equilibrium relatively quickly when migration rate is high, this process can still take tens or hundreds of generations. Furthermore, $F_{ST}$ or $\theta$ by itself cannot distinguish genetic differences that arise due to a migration-drift balance from those that accumulate over time in completely isolated populations. These two scenarios might have very different implications for the concept of what a population is, particularly under the ecological paradigm. Recently developed methods have the potential to distinguish them in some cases (Hey and Nielsen 2004).

On a more technical note, several methods for estimating $\theta$ are available. Although the most commonly used method (and the one used here; Weir and Cockerham 1984) is generally the least biased, other estimators have smaller variance (Weir and Hill 2002). Based on results of computer simulations, Raufaste and Bonhomme (2000) recommended use of Weir and Cockerham’s $\hat{\theta}$ when differentiation is strong but favored a bias-corrected version of Robertson and Hill’s (1984) $\hat{\theta}$ when population subdivision is weak.

Although comparing the number of correct assignments with the random expectation appears to be a powerful method of detecting departures from panmixia, the percentage of correct assignments is not a reliable indicator of the degree of population subdivision. Percent correct assignment is strongly affected by marker type and more weakly by sample size, population size, number of populations, and number of loci (Figure 6; Appendix Table 1). As a consequence, any particular percentage of correct assignments could be consistent with a wide variety of true $Nm$ values.

**Testing migration rate**

Quantitative evaluation of the concept of “population” under the ecological paradigm is challenging for two major reasons. First, the relationship between migration rate and demographic independence is poorly understood. The value $m = 0.1$ for Criterion EC1 is a rough
approximation based on a simple model; real metapopulations will typically be more complex, with population synchrony being a function of both migration rate and correlated environmental fluctuations (Lande et al. 1999). Furthermore, migrant individuals might not be equivalent to local ones in terms of behavior, life history, etc., which means that \( m \) by itself will not necessarily be a reliable indication of the magnitude of demographic interactions.

Second, genetic methods have an inherent difficulty in evaluating the concept of population under the ecological paradigm; demographic independence depends on \( m \), whereas the magnitude of genetic differentiation scales with the product \( Nm \). In part because of this difficulty, recently developed likelihood models that can estimate \( m \) and \( N_e \) separately have attracted a great deal of interest. However, the coalescent approach of Beerli and Felsenstein (2001) has some significant limitations: it is computationally intensive and currently not feasible to use with many typical datasets; it estimates migration rates on an evolutionary time scale that is not directly relevant to the ecological paradigm; and an empirical evaluation (Abdo et al. 2004) indicates that the method performs poorly at estimating migration rates and their confidence intervals. The method of Wang and Whitlock (2003) estimates a contemporary migration rate but requires at least two temporally spaced sets of samples and assumes a migration model that is not realistic for most natural systems. Consequently, although both of these models have the potential to provide important insights into population structure under some circumstances, neither was evaluated in this study.

In some cases, assignment tests also have reasonable power to detect migrant individuals (Paetkau et al. 2004), and in principle this provides a basis for estimating a contemporary migration rate by taking advantage of naturally occurring “genetic marks” of individuals. A limitation of this approach is that the probability of detecting migrants (and hence the estimated migration rate) can depend heavily on the choice of Type I and Type II error rates (Paetkau et al. 2004). This suggests that an assignment method that directly estimates a population-level migration rate might be more powerful and less biased. A Bayesian method to estimate contemporary \( m \) directly was recently proposed by Wilson and Rannala (2003), who also carried out a simulation study in which they considered two populations and a range of migration rates \((m = 0.01-0.20)\) that encompass Criterion EC1 \((m < 0.1)\) for demographic independence. Their results indicate that reliable estimates of \( m \) can be obtained when differentiation is strong \((F_{ST} \sim 0.25)\) and sampling is adequate \((L = 20; S = 100)\), but large biases are observed with insufficient data, particularly for high \( m \) \((F_{ST} = 0.01)\). A more thorough evaluation of Wilson and Rannala’s (2003) method is needed before being able to determine whether it is suitable for estimating migration rates relevant to the ecological paradigm. In particular, it is necessary to further explore the effect of genetic differentiation and investigate the effects of population size, number of actual (and sampled) populations, and the prior distribution for \( \hat{m} \).

Bentzen (1998) suggested one solution to the problem of drawing demographic conclusions from genetic data: he reasoned that if \( m \) is large enough to lead to demographic dependence, \( Nm \) will generally be so large that the genetic signal will be very weak and genetic methods would not be able to reject the hypothesis of panmixia. He argued, therefore, that if genetic data reveal a significant and reproducible difference between populations (no matter how small), this provides strong evidence that the populations are demographically independent. Our results suggest that such a conclusion can be risky; if an adequate number of highly variable
genetic markers is available, genetic structure can be detected consistently even with migration
drates as high or higher ($m = 0.1 – 0.2$) than levels generally thought to lead to correlated
demographic trajectories. For example, in the parameter set Hi-100 ($N = 500$ and $m = 0.2$), $Nm$
was 100 migrants per generation and mean $\hat{\theta}$ was only 0.0014, yet significant population
subdivision ($P<0.05$) was detected over half the time (Appendix Table 1). Based on criterion
EC1 ($m < 0.1$), this would represent a Type I error rate of > 50%. When $N$ is very large,
however, such as marine fish stocks that were the focus of Bentzen’s (1998) evaluations,
migration rates of 10-20% would result in very high $Nm$ values (and even smaller $\hat{\theta}$ ) and hence
a lower Type I error rate under conditions considered here. For very large populations, therefore,
a significant (and repeatable) test of genetic differentiation still might be a reliable indication that
migration is below the threshold for demographic independence—at least until enough highly
variable markers become available to provide arbitrarily high power to detect even smaller
genetic differences.

Limitations of this study

Our ability to conduct in-depth evaluations has been constrained by the huge potential
parameter space and the large number of methods available. Therefore, several limitations of the
current study should be kept in mind in interpreting the results.

First, we considered only a simple island model with constant population sizes and
constant, symmetrical migration, which are unlikely in natural systems. Continuously distributed
species with no apparent population boundaries would present special challenges for any of the
methods described here. Similarly, population structures characterized by isolation by distance
or hierarchical migration patterns could lead to qualitatively different results than are presented
here.

Second, we assumed selective neutrality, in which case the nominal migration rate ($m$) is
also the effective migration rate. In many cases, however, migrants will be at a selective
disadvantage (Nosil et al. 2005) (or, alternatively, at a selective advantage; Ebert et al. 2002)
compared to local individuals. Furthermore, different genes will experience different selective
pressures and hence different rates of effective migration (Rieseberg et al. 1996; Chan and Levin
2005); as a result, measures of genetic differentiation, and results of tests based on population
criteria like those suggested here, might differ depending on which gene loci are surveyed. This
reality argues for careful consideration not only in the choice of population criteria but also in
evaluating results of genetic analyses.

Third, we considered only codominant nuclear loci. Although many standard genetic
analyses such as those described here can be easily modified to accommodate haploid DNA data
from mitochondria or chloroplasts, maternally inherited markers can provide qualitatively
different types of information about population structure. For the ecological paradigm, it is
important to note that recruitment and population growth is contingent on (and typically limited
by) female reproductive success. Because of this reality, Avise (1995) argued that mtDNA data
should be given special consideration in studies of population structure, since evidence for strong
female philopatry implies demographic independence on ecological time frames.
Fourth, the island model used here, and indeed most population genetics models, assumes
discrete generations, which apply to relatively few species. Rannala and Hartigan (1996)
described a method that allows estimation of a gene flow parameter in species with overlapping
generations, but this topic needs additional investigation.

Finally, in non-equilibrium situations, the ecological and evolutionary paradigms can lead
to different conclusions about population structure, for both conceptual and technical reasons.
Are historically panmictic but recently isolated entities populations? Does the answer differ
depending on whether it is viewed from the ecological or the evolutionary paradigm?
Demographic decoupling occurs as soon as immigration stops, whereas genetic measures will
reflect historical connectivity even if no gene flow occurs at present. Therefore, a measure of
contemporary migration rate (based on marked individuals) could potentially detect the
decoupling and provide information relevant to the ecological paradigm, even in the absence of
meaningful genetic differences at the population level.

Summary and future directions

It is apparent from a review of the literature that no consensus has emerged regarding a
quantitative definition of “population.” This is not necessarily a fatal problem; the concept of
“population” is meaningful under each of the paradigms discussed and, potentially, at various
hierarchical levels within each paradigm. It seems reasonable that a variety of criteria could be
appropriate to analyze this diversity of population concepts. We have suggested quantitative
criteria that could be used to define populations under both the evolutionary and ecological
paradigms. The suggested criteria are not exhaustive but might serve as a starting point for
further discussions and evaluations. Results presented here suggest a number of topics that could
form the basis for future research projects. These include:

Assignment tests and population differentiation. It appears that comparing the number of
correct assignments with the random expectation can be a powerful means of detecting
departures from panmixia (if not absolute levels of population differentiation). It would be
useful to compare performance of this method and the multilocus contingency test under a wider
variety of scenarios (especially unbalanced sampling and asymmetrical migration).

Detecting the number of populations. The surprising power of the pairwise contingency
test approach to detect population structure is a good incentive to find a more rigorous solution to
the problem of lack of independence of different pairwise tests. Even after adjusting for multiple
testing, an algorithm is still needed to translate all the pairwise results into inferences about the
number of component gene pools. It seems likely that more sophisticated approaches than the
simple one suggested here will prove to be more robust and powerful.

Methods based on clustering individuals (without a priori information about sample
locations) have limited power when gene flow is moderate or high. We used STRUCTURE as a
representative of this type of analysis, but this is an active area of research and several other
competing programs are available (e.g., Dawson and Belkhir 2001; Corander et al. 2004; Guillot
et al. 2005). Therefore, comparative analyses of these methods are needed. More detailed
evaluations are also needed to better describe parameter spaces that result in high vs low power
for this class of analyses. This is particularly true for nested or hierarchical models of migration,
which Evanno et al. evaluated for low gene flow scenarios. A more thorough evaluation of the
derm performance of Evanno et al.’s Δk method under moderate and high gene flow is also needed.

Ecological paradigm. The ecological population paradigm remains challenging to
analyze using genetic data. Recent theoretical developments offer some promise that this may
derchange in the future if Moore’s Law (computational power doubles every 18 months) continues
to hold and models continue to be refined and made more biologically realistic.

Acknowledgments
We are indebted to Mark Miller, who modified his RXC program to accommodate
multiple datasets and multiple gene loci. We also thank Jérôme Goudet for sharing an
unpublished manuscript, Silvain Piry for providing a version of GENECLASS2 capable of batch
processing many datasets, and Ryan Waples for valuable assistance in generating and analyzing
data used in this report. Jukka Corander, Pip Courbois, Jérôme Goudet, Lorenz Hauser, Mark
Miller, Mary Ruckelshaus, Matthew Stephens, Koen Verhoeven, and an anonymous reviewer
provided useful comments and discussion. Finally, we are grateful to Louis Bernatchez for
encouraging this work.

Author information box
Robin Waples is interested in developing and applying population genetic principles to
real-world problems in ecology, conservation, and management. His research focuses on
population genetics and conservation genetics of marine and anadromous fishes. Oscar
Gaggiotti’s research focuses on developing theory and statistical methods aimed at bridging the
gap between population ecology, population genetics and evolution. Much of his research is
applied to the study of metapopulations.
References


the analysis of genetic population structure. Bioinformatics 20, 2363-2369.
multilocus genotypes to select or exclude populations as origins of individuals. Genetics
degree of population subdivision. Proceedings of the National Academy of Sciences USA 81, 6073-6077.
degree of population subdivision. Proceedings of the National Academy of Sciences USA 81, 6073-6077.
degree of population subdivision. Proceedings of the National Academy of Sciences USA 81, 6073-6077.
degree of population subdivision. Proceedings of the National Academy of Sciences USA 81, 6073-6077.
degree of population subdivision. Proceedings of the National Academy of Sciences USA 81, 6073-6077.


Whitlock MC, McCauley DE (1999) Indirect measures of gene flow and migration: $F_{ST} \neq 1/(4Nm + 1)$. Heredity 82, 1385–70.


Table 1. A representative sampling of definitions of ‘population’ and related terms

<table>
<thead>
<tr>
<th>Population Definitions</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ecological Paradigm</strong></td>
<td></td>
</tr>
<tr>
<td>A group of organisms of the same species occupying a particular space at a particular time</td>
<td>1,2</td>
</tr>
<tr>
<td>A group of individuals of the same species that live together in an area of sufficient size that all requirements for reproduction, survival and migration can be met</td>
<td>3</td>
</tr>
<tr>
<td>A group of organisms occupying a specific geographic area or biome</td>
<td>4</td>
</tr>
<tr>
<td>A set of individuals that live in the same habitat patch and therefore interact with each other</td>
<td>5</td>
</tr>
<tr>
<td>A group of individuals sufficiently isolated that immigration does not substantially affect the population dynamics or extinction risk over a 100-year time frame</td>
<td>6</td>
</tr>
<tr>
<td><strong>Evolutionary Paradigm</strong></td>
<td></td>
</tr>
<tr>
<td>A community of individuals of a sexually reproducing species within which matings take place</td>
<td>7</td>
</tr>
<tr>
<td>A major part of the environment in which selection takes place</td>
<td>8</td>
</tr>
<tr>
<td>A group of interbreeding individuals that exist together in time and space</td>
<td>9</td>
</tr>
<tr>
<td>A group of conspecific organisms that occupy a more or less well defined geographic region and exhibit reproductive continuity from generation to generation</td>
<td>10</td>
</tr>
<tr>
<td>A group of individuals of the same species living close enough together than any member of the group can potentially mate with any other member.</td>
<td>11</td>
</tr>
<tr>
<td><strong>Statistical paradigm</strong></td>
<td></td>
</tr>
<tr>
<td>An aggregate about which we want to draw inference by sampling</td>
<td>12</td>
</tr>
<tr>
<td>The totality of individual observations about which inferences are to be made, existing within a specified sampling area limited in space and time</td>
<td>13</td>
</tr>
<tr>
<td><strong>Variations</strong></td>
<td></td>
</tr>
<tr>
<td>Stock: a species, group, or population of fish that maintains and sustains itself over time in a definable area</td>
<td>14</td>
</tr>
<tr>
<td>Demographic units: those having separate demographic histories</td>
<td>15</td>
</tr>
<tr>
<td>Demes: separate evolutionary units</td>
<td>15</td>
</tr>
<tr>
<td>Interaction group – based on distance an individual might travel during the non-dispersive stage of its life.</td>
<td>16</td>
</tr>
<tr>
<td>Natural population: can only be bounded by natural ecological or genetic barriers</td>
<td>17</td>
</tr>
<tr>
<td>Local population: a) individuals have a chance to interact ecologically and reproductively with other members of the group, and b) some members are likely to emigrate to or immigrate from other local groups</td>
<td>17</td>
</tr>
</tbody>
</table>

References: 1 Krebs (1994); 2 Roughgarden et al. (1989); 3 Huffaker et al. (1984); 4 Lapedes (1978); 5 Hanski and Gilpin (1996); 6 McElhany et al. (2000); 7 Dobzhansky (1970); 8 Williams (1966); 9 Hedrick (2000); 10 Futuyma (1998); 11 Hartl and Clark (1988); 12 Snedecor and Cochrane (1967); 13 Sokal and Rohlf (1969); 14 Booke (1981); 15 Brown and Ehrlich (1980); 16 den Boer (1977, 1979); 17 Andrewartha and Birch (1984).
Table 2. Parameter sets considered in our analyses of the Evolutionary and Ecological paradigms. The following were fixed in all sets: dioecious; random mating; equal sex ratio; finite island model; all subpopulations of constant size $N_e = N$; $K$-allele mutation. Variable input parameters: $n =$ number of subpopulations; $m =$ migration rate; $L =$ number of loci; $S =$ sample size. Diversity data are averages across replicates: $L_P =$ mean number of polymorphic loci; $H_s =$ mean subpopulation gene diversity, calculated over polymorphic loci only.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Input parameters</th>
<th>Diversity</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Set</td>
<td>$n$</td>
<td>$N$</td>
<td>$m$</td>
<td>$Nm$</td>
<td>Mutation</td>
<td>$L$</td>
</tr>
<tr>
<td>Evolutionary</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hi-P</td>
<td>4</td>
<td>500</td>
<td>0.75</td>
<td>375</td>
<td>High</td>
<td>20</td>
</tr>
<tr>
<td>Hi-25</td>
<td>4</td>
<td>500</td>
<td>0.05</td>
<td>25</td>
<td>High</td>
<td>20</td>
</tr>
<tr>
<td>Hi-5</td>
<td>4</td>
<td>500</td>
<td>0.01</td>
<td>5</td>
<td>High</td>
<td>20</td>
</tr>
<tr>
<td>Hi-1</td>
<td>4</td>
<td>500</td>
<td>0.002</td>
<td>1</td>
<td>High</td>
<td>20</td>
</tr>
<tr>
<td>Hi-01</td>
<td>4</td>
<td>500</td>
<td>0.0002</td>
<td>0.1</td>
<td>High</td>
<td>20</td>
</tr>
<tr>
<td>Lo-P</td>
<td>4</td>
<td>500</td>
<td>0.75</td>
<td>375</td>
<td>Low</td>
<td>20</td>
</tr>
<tr>
<td>Lo-25</td>
<td>4</td>
<td>500</td>
<td>0.05</td>
<td>25</td>
<td>Low</td>
<td>20</td>
</tr>
<tr>
<td>Lo-5</td>
<td>4</td>
<td>500</td>
<td>0.01</td>
<td>5</td>
<td>Low</td>
<td>20</td>
</tr>
<tr>
<td>Lo-1</td>
<td>4</td>
<td>500</td>
<td>0.002</td>
<td>1</td>
<td>Low</td>
<td>20</td>
</tr>
<tr>
<td>Lo-01</td>
<td>4</td>
<td>500</td>
<td>0.0002</td>
<td>0.1</td>
<td>Low</td>
<td>20</td>
</tr>
<tr>
<td>100N-25</td>
<td>4</td>
<td>100</td>
<td>0.25</td>
<td>25</td>
<td>High</td>
<td>20</td>
</tr>
<tr>
<td>100N-5</td>
<td>4</td>
<td>100</td>
<td>0.05</td>
<td>5</td>
<td>High</td>
<td>20</td>
</tr>
<tr>
<td>100N-1</td>
<td>4</td>
<td>100</td>
<td>0.01</td>
<td>1</td>
<td>High</td>
<td>20</td>
</tr>
<tr>
<td>2n-25</td>
<td>2</td>
<td>500</td>
<td>0.05</td>
<td>25</td>
<td>High</td>
<td>20</td>
</tr>
<tr>
<td>2n-5</td>
<td>2</td>
<td>500</td>
<td>0.01</td>
<td>5</td>
<td>High</td>
<td>20</td>
</tr>
<tr>
<td>2n-1</td>
<td>2</td>
<td>500</td>
<td>0.002</td>
<td>1</td>
<td>High</td>
<td>20</td>
</tr>
<tr>
<td>8n-25</td>
<td>8</td>
<td>500</td>
<td>0.05</td>
<td>25</td>
<td>High</td>
<td>20</td>
</tr>
<tr>
<td>8n-5</td>
<td>8</td>
<td>500</td>
<td>0.01</td>
<td>5</td>
<td>High</td>
<td>20</td>
</tr>
<tr>
<td>8n-1</td>
<td>8</td>
<td>500</td>
<td>0.002</td>
<td>1</td>
<td>High</td>
<td>20</td>
</tr>
<tr>
<td>10L-25</td>
<td>4</td>
<td>500</td>
<td>0.05</td>
<td>25</td>
<td>High</td>
<td>10</td>
</tr>
<tr>
<td>10L-5</td>
<td>4</td>
<td>500</td>
<td>0.01</td>
<td>5</td>
<td>High</td>
<td>10</td>
</tr>
<tr>
<td>10L-1</td>
<td>4</td>
<td>500</td>
<td>0.002</td>
<td>1</td>
<td>High</td>
<td>10</td>
</tr>
<tr>
<td>25S-25</td>
<td>4</td>
<td>500</td>
<td>0.05</td>
<td>25</td>
<td>High</td>
<td>20</td>
</tr>
<tr>
<td>25S-5</td>
<td>4</td>
<td>500</td>
<td>0.01</td>
<td>5</td>
<td>High</td>
<td>20</td>
</tr>
<tr>
<td>25S-1</td>
<td>4</td>
<td>500</td>
<td>0.002</td>
<td>1</td>
<td>High</td>
<td>20</td>
</tr>
<tr>
<td>C-25</td>
<td>4</td>
<td>500</td>
<td>0.05</td>
<td>25</td>
<td>Low</td>
<td>10</td>
</tr>
<tr>
<td>C-5</td>
<td>4</td>
<td>500</td>
<td>0.01</td>
<td>5</td>
<td>Low</td>
<td>10</td>
</tr>
<tr>
<td>C-1</td>
<td>4</td>
<td>500</td>
<td>0.002</td>
<td>1</td>
<td>Low</td>
<td>10</td>
</tr>
</tbody>
</table>

Ecological

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Input parameters</th>
<th>Diversity</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Hi-100</td>
<td>4</td>
<td>500</td>
<td>0.2</td>
<td>100</td>
<td>High</td>
<td>20</td>
</tr>
<tr>
<td>Hi-50</td>
<td>4</td>
<td>500</td>
<td>0.1</td>
<td>50</td>
<td>High</td>
<td>20</td>
</tr>
<tr>
<td>200N-20</td>
<td>4</td>
<td>200</td>
<td>0.1</td>
<td>20</td>
<td>High</td>
<td>20</td>
</tr>
<tr>
<td>200N-10</td>
<td>4</td>
<td>200</td>
<td>0.05</td>
<td>10</td>
<td>High</td>
<td>20</td>
</tr>
<tr>
<td>200N-2</td>
<td>4</td>
<td>200</td>
<td>0.01</td>
<td>2</td>
<td>High</td>
<td>20</td>
</tr>
<tr>
<td>50N-5</td>
<td>4</td>
<td>50</td>
<td>0.1</td>
<td>5</td>
<td>High</td>
<td>20</td>
</tr>
<tr>
<td>50N-2.5</td>
<td>4</td>
<td>50</td>
<td>0.05</td>
<td>2.5</td>
<td>High</td>
<td>20</td>
</tr>
<tr>
<td>50N-0.5</td>
<td>4</td>
<td>50</td>
<td>0.01</td>
<td>0.5</td>
<td>High</td>
<td>20</td>
</tr>
</tbody>
</table>
Table 3. Number of individuals correctly assigned to population of origin required to demonstrate performance greater than random expectation. It is assumed that each of the $n$ samples includes the same number of individuals ($S$).

<table>
<thead>
<tr>
<th>$n$</th>
<th>$S$</th>
<th>$P&lt;0.05$</th>
<th>$P&lt;0.01$</th>
<th>$P&lt;0.001$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>50</td>
<td>60</td>
<td>65</td>
<td>70</td>
</tr>
<tr>
<td>4</td>
<td>25</td>
<td>32</td>
<td>35</td>
<td>39</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>58</td>
<td>62</td>
<td>65</td>
</tr>
<tr>
<td>8</td>
<td>50</td>
<td>61</td>
<td>66</td>
<td>71</td>
</tr>
</tbody>
</table>
Figure Legends

Figure 1. The continuum of population differentiation. Each group of circles represents a group of subpopulations with varying degrees of connectivity (geographic overlap and/or migration). A) Complete independence. B) Modest connectivity. C) Substantial connectivity. D) Panmixia; “subpopulations” are completely congruent.

Figure 2. Graphical illustration of an ad hoc method of computing the number of different populations represented by a collection of samples. Each circle represents a sample from a potential “population”; dotted lines indicate non-significant results for a multilocus contingency test of heterogeneity of allele frequencies among pairs of samples. Samples that can be linked through a chain of non-significant tests are considered to be part of the same population. In this example, groups of samples A, B, and C represent three different populations.

Figure 3. Power (percent of replicates in which panmixia could be rejected at $P<0.01$) of three methods when true $Nm = 25$. Except as noted, parameters were as in standard model ($N = 500; n = 4; S = 50; L = 20$ High mutation loci). “Combo” = parameter set C-25 (Low mutation, reduced $S$ and $L$).

Figure 4. Power to reject hypothesis that $Nm < 25$ (Criterion EV4) as a function of true $Nm$ and marker type, with other parameters as in the standard model. The hypothesis is rejected if the lower CI for $\hat{\theta}$ is larger than $E(\theta)$ for $Nm = 25$.

Figure 5. Power to reject an hypothesis of restricted gene flow (H0: true $Nm <$ hypothesized $Nm$ at $P<0.05$ level) as a function of true and hypothesized $Nm$. Results (Appendix Table 1 and unpublished data) are for the standard model with $N = 500, n = 4, S = 50$, and $L = 20$ High mutation markers. Dotted line depicts the relationship true $Nm = 0.5 *$ hypothesized $Nm$.

Figure 6. Percent correctly assigned individuals using the classical assignment test (Rannala and Mountain 1997) as a function of the number of migrants per generation ($P =$ panmixia). Except as noted, parameters were as in standard model with High mutation markers. With $n = 4$ subpopulations, the random expectation is 25% correct assignments by chance alone (horizontal dashed line). The diamond symbols connected with a dotted line represent the actual percentage of non-migrants in each population, which sets an upper limit for expected power.

Figure 7. Percent of replicates in which correct number of populations was detected, using three different methods. RXC and BAPS evaluated groups of individuals defined by a priori samples; STRUCTURE performed cluster analysis on individuals. Except as noted, parameters were as in the standard model with $Nm = 5$.

Figure 8. Variation across replicate datasets in number of populations detected, using three different methods. Except as noted, parameters were as in standard model.

Figure 9. Comparison of ability of STRUCTURE and classical assignment tests (Rannala and Mountain 1997) to correctly assign individuals to population of origin. Except as noted, parameters were as in standard model with $Nm = 5$. 
Appendix Figure 1. Estimating the number of populations using STRUCTURE. Each panel shows variation across 10 replicate datasets in ln[P(X|k)] plotted as a function of the putative number of populations (k). For each replicate, results were averaged across five trial runs and scaled to the maximum value within that replicate. The true number of populations was n = 4 and other parameters were as in the standard model; the level of gene flow (N_m) varied as shown in the three panels.

Appendix Figure 2. As in Appendix Figure 1, except that plotted values use the Δk method proposed by Evanno et al. (2005).
Figure 1

A

B

C

D

Isolation

Panmixia

Divergence
Figure 2
Figure 3

Parameter set:
- N = 100
- n = 2
- n = 8
- L = 10
- S = 25
- Low
- Combo

Power

Parameter set

RxC
Assignment
CI (θ)
Figure 4

Actual number of migrants (Nm)

Power

P < 0.05
P < 0.01

High mutation
Low mutation

Actual number of migrants (Nm)
Figure 5

![Graph showing power distribution for hypothesized vs. true Nm values.]

- >85% Power
- <15% Power
- 15-85% Power
Figure 6

The figure shows the relationship between the number of migrants (Nm) and the percent correct assignments. The x-axis represents the number of migrants, ranging from 0 to 0.1, and the y-axis represents the percent correct assignments, ranging from 0 to 100. Different lines represent different conditions: Combo, Standard, L=10, S=25, and Low Mutation. The graph also includes a dashed line indicating the random expectation, with a note pointing to the actual non-migrants.
Figure 7

Parameter set

- Nm=25
- Nm=5
- Nm=1
- n=2
- n=8
- S=25
- L=10
- Low
- Combo

Percent correct

- RxC
- BAPS
- STRUCTURE
Figure 8

- **n = 4; S = 50**
- **n = 4; S = 25**
- **n = 8; S = 50**

### Key
- **RxC**
- **BAPS**
- **STRUCTURE**

### Graph Details
- Populations detected
- Frequency

---

40
Figure 9

![Bar chart showing percent correct assignments for different parameter sets.

Parameter set:
- Nm=25
- Nm=5
- Nm=1
- n=2
- n=8
- S=25
- L=10
- Low
- Combo

Legend:
- STRUCTURE
- Assignment

Y-axis: Percent correct assignments
X-axis: Parameter set

Appendix Table 1. Detailed results of analysis of simulated data, using multilocus contingency tests (RXC), classical assignment tests (Rannala and Mountain 1997) and F-statistics (; Weir and Cockerham 1984). Results reflect data for 100 replicates except for parameter sets Hi-P and Lo-P, for which 1000 replicates were used. Data in bold are empirical Type I error rates for the nominal α level. See Table 2 for input parameters for each parameter set.

<table>
<thead>
<tr>
<th>Param.</th>
<th>Contingency test Percent significant</th>
<th>Percent of replicates # correct &gt; random</th>
<th>Assignment tests</th>
<th>Percent of replicates rejecting H₀ as shown</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Panmixia</td>
</tr>
<tr>
<td>Set</td>
<td>0.05 0.01 0.001</td>
<td>0.05 0.01 0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hi-P</td>
<td>4.9 0.9 0.1</td>
<td>24.6 9.5 3.1 0.5</td>
<td>0.000 0.000</td>
<td>9.3 2.3 0 0 0 0</td>
</tr>
<tr>
<td>Hi-25</td>
<td>100 100 100</td>
<td>48.7 100 100 100</td>
<td>0.006 0.007</td>
<td>100 100 0 0 0 0</td>
</tr>
<tr>
<td>Hi-5</td>
<td>100 100 100</td>
<td>88.7 100 100 100</td>
<td>0.033 0.035</td>
<td>100 100 100 100 0 0</td>
</tr>
<tr>
<td>Hi-1</td>
<td>100 100 100</td>
<td>99.6 100 100 100</td>
<td>0.136 0.136</td>
<td>100 100 100 100 100 2</td>
</tr>
<tr>
<td>Hi-0.1</td>
<td>100 100 100</td>
<td>100.0 100 100 100</td>
<td>0.376 0.395</td>
<td>100 100 100 100 100 100</td>
</tr>
<tr>
<td>Lo-P</td>
<td>5.0 1.0 0</td>
<td>24.9 8.6 1.5 0.3</td>
<td>0.000 0.000</td>
<td>5.6 1.8 0 0 0 0</td>
</tr>
<tr>
<td>Lo-25</td>
<td>79 64 38</td>
<td>32.5 76 51 29</td>
<td>0.007 0.007</td>
<td>60 36 3 0 0 0</td>
</tr>
<tr>
<td>Lo-5</td>
<td>100 100 100</td>
<td>49.8 100 100 100</td>
<td>0.035 0.036</td>
<td>99 99 89 74 2 0</td>
</tr>
<tr>
<td>Lo-1</td>
<td>100 100 100</td>
<td>85.4 100 100 100</td>
<td>0.160 0.158</td>
<td>100 100 100 100 99 2</td>
</tr>
<tr>
<td>Lo-0.1</td>
<td>100 100 100</td>
<td>99.9 100 100 100</td>
<td>0.693 0.652</td>
<td>100 100 100 100 100 100 100</td>
</tr>
<tr>
<td>100N-25</td>
<td>92 80 63</td>
<td>36.2 92 78 65</td>
<td>0.004 0.005</td>
<td>76 63 0 0 0 0</td>
</tr>
<tr>
<td>100N-5</td>
<td>100 100 100</td>
<td>71.5 100 100 100</td>
<td>0.032 0.036</td>
<td>100 100 100 100 3 0 0</td>
</tr>
<tr>
<td>100N-1</td>
<td>100 100 100</td>
<td>96.7 100 100 100</td>
<td>0.150 0.153</td>
<td>100 100 100 100 100 2 0</td>
</tr>
<tr>
<td>2n-5</td>
<td>82 58 34</td>
<td>64.2 85 72 55</td>
<td>0.004 0.006</td>
<td>49 26 1 0 0 0</td>
</tr>
<tr>
<td>2n-1</td>
<td>100 100 100</td>
<td>87.1 100 100 100</td>
<td>0.023 0.024</td>
<td>100 100 100 94 1 0 0</td>
</tr>
<tr>
<td>2n-5</td>
<td>100 100 100</td>
<td>99.3 100 100 100</td>
<td>0.097 0.100</td>
<td>100 100 100 100 100 1 0</td>
</tr>
<tr>
<td>8n-5</td>
<td>100 100 100</td>
<td>39.2 100 100 100</td>
<td>0.008 0.009</td>
<td>100 100 0 0 0 0</td>
</tr>
<tr>
<td>8n-5</td>
<td>100 100 100</td>
<td>89.3 100 100 100</td>
<td>0.039 0.040</td>
<td>100 100 100 100 2 1 0 0</td>
</tr>
<tr>
<td>8n-5</td>
<td>100 100 100</td>
<td>99.6 100 100 100</td>
<td>0.147 0.152</td>
<td>100 100 100 100 1 1 0 0</td>
</tr>
<tr>
<td>10L-25</td>
<td>100 100 100</td>
<td>42.4 100 100 98</td>
<td>0.007 0.007</td>
<td>99 92 2 1 0 0 0 0</td>
</tr>
<tr>
<td>10L-5</td>
<td>100 100 100</td>
<td>77.2 100 100 100</td>
<td>0.034 0.035</td>
<td>100 100 100 100 1 1 0 0</td>
</tr>
<tr>
<td>10L-1</td>
<td>100 100 100</td>
<td>98.3 100 100 100</td>
<td>0.137 0.136</td>
<td>100 100 100 100 100 5 1</td>
</tr>
<tr>
<td>25S-25</td>
<td>98 93 67</td>
<td>43.4 98 94 74</td>
<td>0.007 0.007</td>
<td>89 72 1 0 0 0 0 0</td>
</tr>
<tr>
<td>25S-5</td>
<td>100 100 100</td>
<td>84.5 100 100 100</td>
<td>0.034 0.035</td>
<td>100 100 100 100 5 0 0 0</td>
</tr>
<tr>
<td>25S-1</td>
<td>100 100 100</td>
<td>99.5 100 100 100</td>
<td>0.136 0.136</td>
<td>100 100 100 100 3 2</td>
</tr>
<tr>
<td>C-25</td>
<td>18 5 1</td>
<td>27.4 24 13 3</td>
<td>0.005 0.006</td>
<td>10 7 1 0 0 0 0 0</td>
</tr>
<tr>
<td>C-5</td>
<td>91 86 69</td>
<td>40.3 91 82 60</td>
<td>0.034 0.036</td>
<td>71 52 46 27 4 2 0 0</td>
</tr>
</tbody>
</table>
Appendix Table 1, continued

<table>
<thead>
<tr>
<th>Param.</th>
<th>Percent significant</th>
<th>Contingency test</th>
<th>Percent of replicates # correct &gt; random</th>
<th>Assignment tests</th>
<th>Percent of replicates rejecting $H_0$ as shown</th>
<th>$\hat{\theta}$</th>
<th>$E(\hat{\theta})$</th>
<th>Panmixia</th>
<th>$Nm \geq 25$</th>
<th>$Nm \geq 5$</th>
<th>$Nm \geq 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hi-100</td>
<td>0.05 0.01 0.001</td>
<td>56 33 15</td>
<td>31.0 55 39 17</td>
<td>0.0014 0.0019</td>
<td>45 24</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
</tr>
<tr>
<td>Hi-50</td>
<td>0.05 0.01 0.001</td>
<td>98 92 77</td>
<td>37.5 98 88 71</td>
<td>0.003 0.004</td>
<td>92 85</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
</tr>
<tr>
<td>200N-20</td>
<td>0.05 0.01 0.001</td>
<td>99 99 99</td>
<td>46.9 99 99 99</td>
<td>0.008 0.009</td>
<td>99 98</td>
<td>1 1</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
</tr>
<tr>
<td>200N-10</td>
<td>0.05 0.01 0.001</td>
<td>100 100 100</td>
<td>63.5 100 100 100</td>
<td>0.016 0.018</td>
<td>100 100</td>
<td>63 47</td>
<td>1 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
</tr>
<tr>
<td>200N-2</td>
<td>0.05 0.01 0.001</td>
<td>100 100 100</td>
<td>95.2 100 100 100</td>
<td>0.080 0.083</td>
<td>100 100</td>
<td>100 100</td>
<td>100 100</td>
<td>100 100</td>
<td>100 100</td>
<td>100 100</td>
<td>1 0</td>
</tr>
<tr>
<td>50N-5</td>
<td>0.05 0.01 0.001</td>
<td>100 100 100</td>
<td>58.9 100 100 100</td>
<td>0.029 0.036</td>
<td>100 100</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
</tr>
<tr>
<td>50N-2.5</td>
<td>0.05 0.01 0.001</td>
<td>100 100 100</td>
<td>74.6 100 100 100</td>
<td>0.063 0.069</td>
<td>100 100</td>
<td>60 40</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
<td>0 0</td>
</tr>
<tr>
<td>50N-0.5</td>
<td>0.05 0.01 0.001</td>
<td>100 100 100</td>
<td>96.9 100 100 100</td>
<td>0.267 0.265</td>
<td>100 100</td>
<td>100 100</td>
<td>100 100</td>
<td>100 100</td>
<td>100 100</td>
<td>100 100</td>
<td>4 2</td>
</tr>
</tbody>
</table>
Appendix Figure 2

Scaled $\Delta(k)$

$Nm = 1$

$Nm = 5$

$Nm = 25$

Number of populations