Molecular Ecology (2004) **13**, 55–65 doi: 10.1046/j.1365-294X.2003.02008.x

Genetic assignment methods for the direct, real-time estimation of migration rate: a simulation-based exploration of accuracy and power

DAVID PAETKAU,*† ROBERT SLADE,‡§ MICHAEL BURDEN§ and ARNAUD ESTOUP*¶
Department of Zoology and Entomology, University of Queensland, St. Lucia, QLD 4072, Australia, †Wildlife Genetics International, Box 274, Nelson, BC V1L 5P9, Canada, ‡Graduate Research College, Southern Cross University, Australia, §Australian National Genomic Information Service, University of Sydney & NSW 2006, Australia, ¶INRA, Centre de Biologie et de Gestion des Populations, Campus International de Baillarguet, CS 30 016, 34900, Monferrier/Lez cedex, France

Abstract

Genetic assignment methods use genotype likelihoods to draw inference about where individuals were or were not born, potentially allowing direct, real-time estimates of dispersal. We used simulated data sets to test the power and accuracy of Monte Carlo resampling methods in generating statistical thresholds for identifying F_0 immigrants in populations with ongoing gene flow, and hence for providing direct, real-time estimates of migration rates. The identification of accurate critical values required that resampling methods preserved the linkage disequilibrium deriving from recent generations of immigrants and reflected the sampling variance present in the data set being analysed. A novel Monte Carlo resampling method taking into account these aspects was proposed and its efficiency was evaluated. Power and error were relatively insensitive to the frequency assumed for missing alleles. Power to identify F_0 immigrants was improved by using large sample size (up to about 50 individuals) and by sampling all populations from which migrants may have originated. A combination of plotting genotype likelihoods and calculating mean genotype likelihood ratios (D_{LR}) appeared to be an effective way to predict whether F_0 immigrants could be identified for a particular pair of populations using a given set of markers.

Keywords: admixture linkage disequilibrium, genetic assignment, immigrants, migration rate, power, statistical significance

Received 13 June 2003; revision received 9 September 2003; accepted 9 September 2003

Introduction

Understanding connectivity between populations is a fundamental aspect of population ecology (Wiens 1997), and an essential component of managing populations for conservation. Geneticists have taken an interest in this subject ever since Wright (1931) demonstrated that indirect estimates of *Nm* (the effective number of individuals moving between populations per generation) could be generated from measures of genetic differentiation between populations. The limitations of such indirect approaches are that they are based on simplified, and typically unrealistic, population models, and that there can be high variance associated with the estimates they produce. Noting these limitations, Whitlock & McCauley (1999) recently concluded that estimates of *Nm* derived from F_{ST} were only 'likely to be correct within a few orders of magnitude'; a degree of precision of little value for understanding demographic processes in real populations. Furthermore, indirect estimates of *Nm* reflect long-term dispersal rates, and will be slow to respond to the 'real-time' changes in movement patterns across specific landscapes that we need to be able to measure for management purposes.

Population genetic analyses, like the estimation of F_{ST}, have traditionally been based on the comparison of allele frequency distributions observed in population samples. However, with the development of highly variable genetic markers, like microsatellites, it is now possible to collect enough information about individual organisms for the units of analysis to be shifted from populations to individuals

Correspondence: A. Estoup, INRA-CBGP, Campus International de Baillarguet, CS 30 016, 34900, Monferrier/Lez cedex, France. E-mail: estoup@ensam.inra.fr

(for review see Estoup & Angers 1998). With this shift, it has become possible for genetic analysis to yield information on population ecology that is qualitatively similar to that obtained with traditional field techniques (e.g. Woods *et al*. 1999). This has encouraged the pursuit of individual-level genetic methods for studying population of origin (Paetkau *et al*. 1995; Rannala & Mountain 1997; Waser & Strobeck 1998; Cornuet *et al*. 1999; Davies *et al*. 1999; Pritchard *et al*. 2000; Eldridge *et al*. 2001; Guinand *et al*. 2002; Maudet *et al*. 2002; Wilson & Rannala 2003).

The hope for using genetics to directly study interpopulation movements lies in the ability to identify immigrants (Rannala & Mountain 1997; Pritchard *et al*. 2000; Wilson & Rannala 2003), or even individuals' natal populations (Cornuet *et al*. 1999). In many settings, what one really wants to do is to make a direct, real-time estimation of a migration rate in a population system with ongoing geneflow. In this context, the null hypothesis is that an individual was born in the population in which it was sampled (including individuals born to immigrant parents, i.e. F_1, F_2, etc. immigrants). Paetkau *et al*. (1995) demonstrated that it was often possible to assign individuals to the population where they were sampled using the probability of drawing that individual's multilocus microsatellite genotype from the allele distributions observed in a series of study populations. One major challenge is to distinguish between residents (individuals born in the population in which they are sampled) that are 'misassigned' (have a genotype that is most likely to occur in a population other than the one in which the individual was sampled) by chance and F_0 immigrants who are misassigned because they were born somewhere other than where they were sampled. A potential solution to this problem is to use Monte Carlo resampling methods to identify statistical thresholds beyond which individuals are likely to be F_0 immigrants (Rannala & Mountain 1997; Cornuet *et al*. 1999; we use 'sampling' to refer to the collection of samples from populations, and 'resampling' to refer to the drawing of genotypes from that population data). The principle behind this approach is to approximate the distribution of genotype likelihoods that would be found in a particular group of 'real' resident individuals, and then to compare the likelihoods calculated for sampled individuals to that distribution. Wilson & Rannala (2003) recently proposed a Bayesian method relying on Markov chain Monte Carlo techniques that attempts to estimate posterior probabilities of a number of population parameters, including recent migration rates (over several generations); it is worth emphasizing that our interest was much narrower, focusing solely on the identification of immigrant individuals in the current generation.

In the present paper, we explore the Monte Carlo resampling methods of Rannala & Mountain (1997) and Cornuet *et al*. (1999), as well as a novel Monte Carlo resampling method, through the use of simulated data in which F_0 immigrants and residents are known. In particular, we were concerned that these methods might not generate realistic likelihood distributions, and that this might lead to an excess of resident individuals being identified as F_0 immigrants. We were particularly keen to know whether this approach was sufficiently powerful to be of practical value in studying dispersal and estimating migration rates between populations having reasonably high historic and ongoing gene flow. In other words, are there situations where Nm is high enough that a realistic population sample would contain enough immigrants to shed light on immigration patterns, yet where there remained enough differentiation between populations to endow genetic assignment methods with adequate power for F_0 immigrant detection. The aim of our simulation study differs from previous simulation studies that explored the efficiency of the Monte Carlo resampling methods (Rannala & Mountain 1997; Cornuet *et al*. 1999). In particular, the simulations described in Cornuet *et al*. (1999) used a model in which populations diverged from a common ancestral population without migration, and hence did not focus on the specific task of immigrant identification.

Although one can imagine many scenarios and questions that would be interesting to explore, our study employed an island population model, and focused on the following questions: (i) does the Monte Carlo resampling procedures produce more type I errors (we define a type I error as a resident individual falling beyond a critical value α, and thus being wrongly identified as an F_0 immigrant) than appropriate given the specified α and, if so, can a better procedure be found? (ii) what is the impact of assuming different frequencies for dealing with missing alleles? (iii) how do sampling parameters (number of markers, sample size per population, number of populations sampled) affect power (we define power as the proportion of true F_0 immigrants falling past a chosen critical value; power = 1 – type II error)? (iv) what is the best way to predict power in empirical data sets? Reasoning that the power and practical utility of these methods would be highest in the sort of small, isolated populations that are of high relevance to conservation, we selected our simulation parameters to reflect such populations.

Methods

Simulation methods (data generation)

Genotypic data were generated using a simulation method that involved a backward process followed by a forward process. The backward process was based on the coalescence theory (Hudson 1991). Coalescent trees were simulated using a 'generation by generation' algorithm (e.g. Leblois *et al*. 2003), which allowed great flexibility in the choice, and variation over time, of demographic parameters,

including migration rate (m, the proportion of genes migrating between populations per generation) and the effective number of diploid individuals in the population (N). With this backward process, single-locus data sets were produced independently and combined to produce a single data set of statistically unlinked loci.

In the backward process, migration was a probabilistic procedure that occurred independently for each locus, and immigrant genes were distributed randomly across the final multilocus genotypes. This stands in contrast to the process in nature where immigrant genes arrive in a population as a set, in an immigrant individual or gamete. This association of immigrant genes (i.e. admixture linkage disequilibrium; Stephens *et al*. 1994) in immigrants and their descendents may have an affect on the power to differentiate between F_0 immigrants and residents (that include the descendents of immigrants from previous generations). Since non–random associations of immigrant genes would extend the tail of genotype likelihood distributions, we were concerned that a strict backward coalescent approach, in which associations are random, might produce exaggerated estimates of power.

If loci are unlinked, admixture linkage disequilibrium decays by half each generation, so a reasonably realistic level of admixture linkage disequilibrium could be produced by introducing the disequilibrium resulting from the few most recent generations of immigrants. To do this, we followed the backward process by 10 generations of a classical forward simulation process wherein generations were produced by drawing multilocus haploid gametes, with replacement, from randomly chosen individuals in the previous generation. In this process, immigration involved the probabilistic movement of multilocus diploid genotypes (i.e. whole individuals).

It is worth noting that the forward process made it necessary to simulate $2N$ genes during the backward process, whereas coalescence methods normally only require the simulation of as many genes as are sampled. The usual limitation of small sample size compared to $2N$ of many coalescence simulation methods did not hold here. As a matter of fact, the generation by generation algorithm used for our backward process does not approximate coalescence times and allow multiple coalescence events to occur at each generation. However, the necessity to simulate $2N$ genes during the backward process placed relatively tight constraints on our simulation procedure in terms of computing time. As a result, our set of populations never included more than 2000 diploid individuals. The backward, forward and backward-forward processes were tested by comparing analytical expectations of gene diversity (H; Nei & Roychoudhury 1974), and F_{ST} (Weir & Cockerham 1984) to values observed in simulated data sets. Results were indistinguishable from expectation for all processes (not shown).

In the final (sampled) generation of the forward process, exactly Nm individuals were moved into each population, rather than allowing probabilistic migration. This simplified the measurement of error rates, as the analysis program (below) could be told how many immigrants were in each population. Where the product of m and sample size (s; the number of individuals sampled per population) was not a whole number, the number of immigrants moved in the final generation was increased to the nearest whole number. Note that this increased the migration rate above equilibrium levels for the final generation, but should have introduced a conservative bias since it will have caused the populations to be less differentiated than expected at equilibrium, reducing the distinction between residents and immigrants, and thus power.

The parameters varied in this project included m, N, s, and the number of loci (l). It was our intention to loosely approximate typical, contemporary, microsatellite data sets in our simulations, so the mutation model was assumed to be a stepwise mutation model (SMM; Ohta & Kimura 1973), the mutation rate (μ) was high (as expected for loci with a substantial level of polymorphism in populations of relatively small size, i.e. usually 5×10^{-3}) and the number of genotyped loci (l) was 10 except where stated otherwise. Each data set contained 10 populations of equal size connected in an island model (Wright 1931), and 100 data sets were produced for each combination of parameters explored. Means and coefficients of variation were calculated for each group of 100 data sets for several population genetic statistics including H, F_{ST}, D_S (Nei 1972) $(\delta\mu)^2$ (Goldstein *et al*. 1995) and D_{LR} (Paetkau *et al*. 1997).

Data analysis

Genotype likelihoods were calculated for each individual in each population following Paetkau *et al*. (1995). Although other assignment criterions may be computed (e.g. Rannala & Mountain 1997; Cornuet *et al*. 1999), the criterion of Paetkau *et al*. (1995) is computationally simple and is the most frequently employed in empirical studies (reviewed in Guinand *et al*. 2002). Moreover, assignment based on the latest likelihood estimator was found to compare favourably with other more sophisticated assignment methods based on artificial neural networks (Guinand *et al*. 2002).

The first step in our assignment process was to calculate observed frequencies for each allele in each population. Then the likelihood of a diploid genotype occurring in a particular population was estimated, under an assumption of random mating, as the square of the observed allele frequency for homozygotes or twice the product of the two allele frequencies for heterozygotes. Likelihoods for each locus were multiplied together, under an assumption of independence between loci, to yield an overall likelihood

for a given individual's genotype in a given reference population. Where an individual was included in the sample for a particular population, that individual was removed from the sample prior to calculating observed allele frequencies (the leave-one-out procedure; Efron 1983; Paetkau *et al*. 1998).

When the observed frequency of any allele in a given genotype was zero (a missing allele), this calculation produced a multilocus likelihood of zero. To avoid such zero values, missing alleles were given an arbitrary nonzero frequency. Six values were tested for missing alleles: 2×10^{-2}, 1×10^{-2}, 5×10^{-3}, 1×10^{-3}, 1×10^{-4} and 1×10^{-5}. Except when testing the different ways of treating missing alleles, simple observed values were used for allele frequencies and missing alleles were given a frequency of 1×10^{-2}.

All individuals were ranked using the test statistic $\Lambda = L_h/L_{max}$, where L_h is the likelihood of drawing that individual's genotype from the population in which it was sampled, given the observed set of allele frequencies, and L_{max} is the maximum such likelihood observed for this genotype in any population (i.e. the likelihood for the population to which the individual would be assigned in the assignment test; Paetkau *et al*. 1995). The test statistic Λ is appropriate when all source populations for immigrants are thought to be sampled. When some source populations are clearly missing, it becomes more appropriate to use L_h as the test statistic. L_h is essentially the assignment criterion previously used by Cornuet *et al*. (1999), as well as by Favre *et al*. (1997) to test for sex-biased dispersal. When exploring the effect on power of failing to sample all source populations, we calculated both L_h and Λ; otherwise Λ was used as the test statistic.

To generate critical values for rejecting the null hypothesis (i.e. that an individual was born in the population in which it is sampled), diploid genotypes were resampled from each of the 10 populations. This was done in two different ways: (i) a classical Monte Carlo resampling method (e.g. Rannala & Mountain 1997; Cornuet *et al*. 1999) where a simulated individual is obtained by drawing alleles at random according to the allele frequencies observed at each locus in each population of the data set; and (ii) a new Monte Carlo resampling method where a simulated individual is obtained by the drawing, with replacement, of multilocus gametes from randomly chosen individuals in each population of the data set. The classical and new procedures were reiterated until the required number of simulated individuals was reached for each population. In the new resampling procedure, one individual (i.e. a potential 'parent') is first randomly drawn in the 'reference' population sample, and one gene and corresponding allelic state is then randomly drawn among the two gene copies for each locus. A second gamete is designed the same way from a second individual ('parent') drawn at random, and both gametes are associated to give a simulated multilocus diploid genotype. Drawing gametes instead of alleles has the effect of retaining the admixture linkage disequilibrium present in sampled immigrants and their recent descendents, with the expected decay of 0.5 between the sample data and the resampled genotypes. Essentially, this sampling procedure is truer to the process that occurs in a natural population when gametes are drawn from one generation to produce the next generation's resident individuals, whose likelihood distribution we were trying to approximate. In this paper, resampling was based on gametes except when specifically testing the effect of sampling alleles instead of gametes.

The test statistics were calculated from the resampled genotypes in two different ways. The first approach was to generate 500 resampled individuals (an arbitrary value much greater than s) for each population, calculate allele frequencies using those large sets of samples, and then calculate genotype likelihoods. This was repeated until a total of 10 000 genotypes (i.e. 20 sets) had been resampled from each population sample. This approach is similar to that of Rannala & Mountain (1997) and Cornuet *et al*. (1999), who calculated test statistics on a single set of 1000 resampled genotypes. Our alternate approach, which was used for all runs except when comparing these two resampling methods, was to calculate allele frequencies and genotype likelihoods based on sets containing exactly as many genotypes as did the original data set (50 per population in the case of the direct comparison between methods), and then repeat the process until the total number of resampled genotypes for which likelihoods had been calculated was very large (exactly 10 000 in the case of direct comparison). These two methods of calculating likelihoods for resampled genotypes differed only in that the allele frequency estimates in the first approach will be more precise than will be the case for the second approach, where sampling error may cause considerable differences in allele frequencies between the sample (in this case simulated) population data and the resampled populations produced from that data. In both cases, likelihoods for all of the genotypes generated by the Monte Carlo procedure (i.e. 10 000 per population) were ranked using the relevant test statistic, and critical values were identified as the value of the test statistic beyond which the proportion of resampled genotypes with equal or more extreme (smaller) values was α.

In the described procedure, we were essentially testing every individual for its status as resident or F_0 immigrant. Our simulated sample data sets contained between 120 and 960 individuals (12–96 per population), with up to 950 residents (when $Nm = 1$). Assuming that type I errors rates were consistent with α, and setting α to 0.05, we would expect a mean of $0.05 * 950 = 47.5$ type I errors in such a data set. Given that only 10 immigrants would be present in this example, this rate of error is clearly unacceptable for most applications, even if all F_0 immigrants also fall

beyond the critical value (i.e. power = 1). One is always faced with a compromise between α and power, and in this study we considered that 0.01 or 0.002 represented appropriate values of α for testing the practical utility of the described methods.

In our simulated sample sets, F_0 immigrants and residents were known. Hence, For each data set, the number of residents and F_0 immigrants falling past the critical value could be recorded, and mean error rates were calculated for the 100 data sets analysed under a given set of parameters and analysis conditions. Power was calculated as [number-of-F_0-immigrants-past-critical-value/total-number-of-F_0-immigrants], and the rate of type I errors relative to expectation was taken as [number-of-residents-falling-past-critical-value/(10 populations * N * $(1 - m)$ * α)]. Note that we based this latter calculation on the known number of simulated residents, whereas in an applied situation one would base expected type I error on the total sample size, because the number of residents and immigrants would be unknown.

In practice, a direct and simple estimation of a migration rate could be obtained by dividing [the total number of individuals falling past the critical value minus the number of expected type I errors] by [the total number of sampled individuals]. However, our simulated data were restricted to simple patterns of balanced and, in the final generation, deterministic migration. Therefore, such migration rate estimations are not presented, and the performance of these methods in more realistic scenarios of asymmetric and stochastic migration is a subject that is open to further study.

The described Monte Carlo resampling methods have been incorporated in the computer program GENECLASS 2, which can be found at http://www.montpellier.inra.fr/CBGP/softwares/.

Results and Discussion

Resampling procedure

In the method that we were studying, the goal of the Monte Carlo resampling procedure was to produce a distribution of a test statistic that reflected the distribution expected for native-born (resident) animals from the sampled population. If the distribution produced is inaccurate, the proportion of type I errors might be different than expected (α). There were two aspects of previous resampling procedures (Rannala & Mountain 1997; Cornuet *et al*. 1999) that we thought might cause higher than expected type I error rates.

First, when analysing a single very large set of resampled genotypes (e.g. Rannala & Mountain 1997; Cornuet *et al*. 1999), or several sets that are larger than the sample data file, the sampling variance for allele frequencies in the simulated data set(s) is expected to be lower than in the actual data set. This may result in a distribution of genotype likelihoods characterized by a reduced tail compared to the actual distribution for real, resident animals, which we are attempting to approximate. Hence, analysing single or several sets of resampled genotypes that are larger than the actual data set, may result in an excess of resident individuals being rejected and wrongly considered as F_0 immigrants (type I errors).

Second, by resampling alleles instead of gametes, recent immigrant alleles become randomly distributed across genotypes instead of being concentrated in individuals descended from recent immigrants (admixture linkage disequilibrium), as expected in populations with ongoing geneflow.

For instance, consider a population of 50 individuals at equilibrium for $Nm = 10$. If we looked at individuals born in this population, approximately one third will be F_1s, having had one parent that was an immigrant in the previous generation, and we can expect two of 50 newborn resident individuals to actually be the offspring of two immigrant parents. Since F_1s have 50% immigrant genes, and the offspring of two immigrants have exclusively immigrant genes, the distribution of genotype likelihoods in this population will have an extended tail of rare genotypes compared to the distribution we would see if we distributed the immigrant genes randomly across all genotypes. Given that the critical values will, by definition, lie in this tail, we would expect that sampling alleles would result in an excess of resident individuals (especially those with recent immigrant ancestry) being rejected as immigrants, and wrongly considered as F_0 immigrants (type I errors). This argument applies primarily to populations with small N and large m; the type of populations we focused on.

To explore these two aspects of the resampling procedure, we used simulations with $N = s = 50$, $\mu = 0.005$, $l = 10$, $Nm = \{0.25, 0.5, 1, 2, 3, 4, 5, 6, 7, 15\}$ (1000 data sets total) and $\alpha = 0.002$. These data files were analysed in four different ways, with each combination of the two factors of interest (i.e. re-sampled populations equal to, vs. 10 times larger than, s; sampling gametes vs. alleles).

Our results indicated that resampling methods can have a large impact (Fig. 1). When resampling was performed on alleles instead of gametes, the proportion of type I error was much higher than α, especially at low values of Nm. Similarly, excessive type I error was observed when re-sampled individuals were analysed in data sets 10 times larger than the original. The combination of these two problems lead to type I error rates seven to eight times higher than expected.

We were not able to explore all of the methods that have been suggested for genetic assignment of individuals to populations, especially the Bayesian methods of Rannala & Mountain (1997). However, we believe that our findings

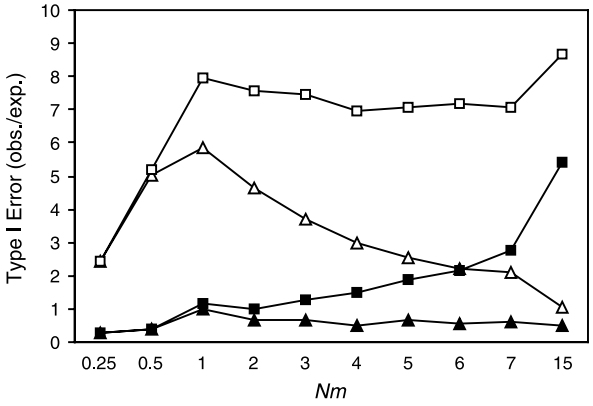

Fig. 1 Four different methods for drawing genotypes from the sample data to produce distributions of the test statistic A: drawing gametes (filled symbols) vs. alleles (open symbols) and analysing genotypes in sets of 10 times (squares) or 1 times (triangles) the size of the original data set. $N = s = 50$, $\mu = 0.005$, $l = 10$, $\alpha = 0.002$.

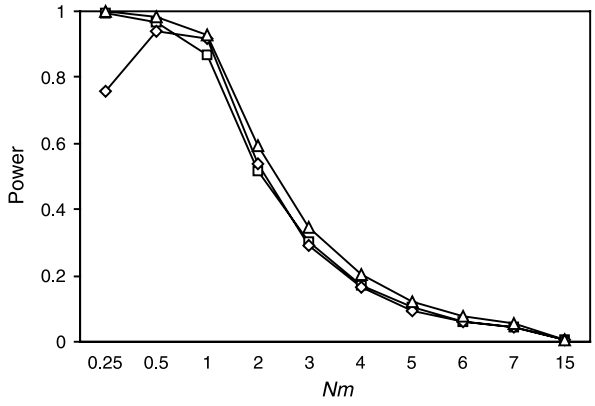

Fig. 2 Power observed when arbitrary frequencies of 2×10^{-2} (squares), 5×10^{-3} (triangles) or 1×10^{-4} (diamonds) were substituted for alleles not observed in a reference sample (missing alleles). Simulated sample files as in Fig. 1. Curves for 1×10^{-2} and 1×10^{-3} were almost identical to the curve for 5×10^{-3}, so were removed for clarity.

have important implications for the use of assignment methods to make direct, real-time estimates of migration rate, that extend beyond the specific methods that we tested (based on the assignment criterion of Paetkau *et al*. 1995). This was illustrated by analysing three randomly chosen simulated data sets using the assignment program IMMANC (Rannala & Mountain 1997) and setting this program to consider only the current generation (as we are doing in this paper). For data sets containing one true F_0 immigrant per population ($Nm = 1$), both programs correctly identified all 30 (10×3) F_0 immigrants. However, while our program ($\alpha = 0.01$) identified 16 residents as F_0 immigrants (mean type I error = 16/(3*490) = 0.0109), Immanc identified 113 residents as immigrants at a significance level (α) of 0.01 (mean type I error = 0.078); a nearly eight-fold excess of type I error compared to expectation, in keeping with our findings regarding resampling methods. We then implemented the assignment criterion of Rannala & Mountain (1997) (cf. Formula 9 in their paper) in our program; our new Monte Carlo resampling method applied with this assignment criterion allowed all 30 F_0 immigrants to be correctly identified and reduced the number of residents identified as F_0 immigrants to 17 (mean type I error = 0.0116).

A different approach based on a Bayesian method relying on Markov chain Monte Carlo techniques was recently proposed by Wilson & Rannala (2003) to estimate posterior probabilities of a number of population parameters, including recent migration rates over the last several generations, among populations. It would be worth comparing in further studies the performance of the latest method with the approach we propose in the present paper.

Missing Alleles

The problem of missing alleles arises when one or more alleles in an individual's genotype has an observed frequency of zero in a reference sample being used for likelihood calculation; the calculated likelihood will also be zero. When Paetkau *et al*. (1995) introduced the use of genotype likelihoods to assign individuals to populations, they avoided missing alleles by adding an individual's genotype to any reference allele distribution in which that sample was not already included. However, this approach was found to reduce the power to correctly assign individuals to populations, and a leave-one-out procedure was subsequently suggested in which individuals were removed from distributions in which they were included, and a value of 0.01 was substituted wherever alleles were missing (Paetkau *et al*. 1998). This solution has the disadvantage that the value of 0.01 is arbitrary.

We investigated the impact of assuming different frequencies for missing alleles using the same simulated data files as in the previous section and $\alpha = 0.002$. Fixed missing-allele frequency values of 2×10^{-2}, 1×10^{-2} ($1/2N$), 5×10^{-3}, 1×10^{-3}, 1×10^{-4} and 1×10^{-5} were used. For all Nm values studied, results were not strongly affected by different frequencies assumed for missing alleles (Fig. 2). While a small, but noticeable, decrease in power was observed for extreme fixed values of 2×10^{-2} and 1×10^{-4}, there was little difference in results when values in the broad range of $1 \times 10^{-2} - 1 \times 10^{-3}$ were used.

An interesting difference between very large (2×10^{-2}) and very small (1×10^{-4}) fixed values is that the latter showed a sharp drop in power at the smallest Nm value that we used (Fig. 2). This spike was even more exaggerated

with a fixed value of 1×10^{-5}, and was even present, although very small, with a value of 1×10^{-3}. An explanation for this increase in type II error is that the extremity of the missing allele correction makes Λ excessively sensitive to the number of missing alleles present in a given multilocus genotype. In other words, a genotype that is otherwise relatively common in a given population would have its calculated likelihood reduced by several orders of magnitude if it contained a single missing allele, dramatically altering its position relative to the calculated critical value. Davies *et al*. (1999) suggested that fixed values as low as 1×10^{-6} could be used for missing alleles, but our testing shows that this extreme value could reduce power considerably, at least under some circumstances.

The optimal missing allele correction for our data sets was a fixed frequency of 5×10^{-3}, but the superiority of this value was subtle and may not hold for situations where the population and sample parameters are much different than ours. The important implication of our results is that the lack of a precise analytical correction for missing alleles is not a significant problem because results are relatively insensitive to the frequency assumed for those alleles.

Sampling parameters

To this point, we have assumed that every individual in every population from which immigrants may have derived has been sampled. In this section, we considered what happens when these assumptions are not met, as well as the impact of the number of genotyped markers (l) and the number of sampled individuals (s). The markers that we used were uniform in their evolutionary characters (i.e. same mutation model and rate, and evolutionary neutrality), whereas empirical (i.e. nonsimulated) studies would make use of markers that vary more in those characters and hence in their specific contribution to population assignment and immigrant detection. Therefore, the question of determining which combination of loci would most likely provide a specific level of population assignment and immigrant detection was not addressed here, and would deserve specific investigation, preferably performed on real (i.e. nonsimulated) microsatellite data (cf. Banks *et al*. 2003).

If samples are not available from all source populations for immigrants, or if only a single population has been sampled, then L_h should be used as a test statistic (e.g. Favre *et al*. 1997). We explored the effect of test statistic on power by analysing the previous data sets using L_h or Λ, at three different critical values ($\alpha = 0.05, 0.01, 0.002$). When the effect of the sample size (s) was assessed, we kept the proportion of F_0 immigrants in the sampled individuals constant. Therefore, we set Nm at 8, and sampled between 1/8th ($s = 12$) to 8/8ths ($s = 96$) of the population ($N = 96$), yielding data sets with 1–8 immigrants in them, but always with 1/12th of sampled individuals being F_0 immigrants. Due to the relatively low power at $Nm = 8$, we used 15 highly variable ($\mu = 0.005$, $H = 0.825$) markers. To study the effect of l on power, we varied l from 5 to 20 while holding other parameters constant ($N = s = 50$, $\mu = 0.0025$, $Nm = \{1, 3, 5, 7\}$).

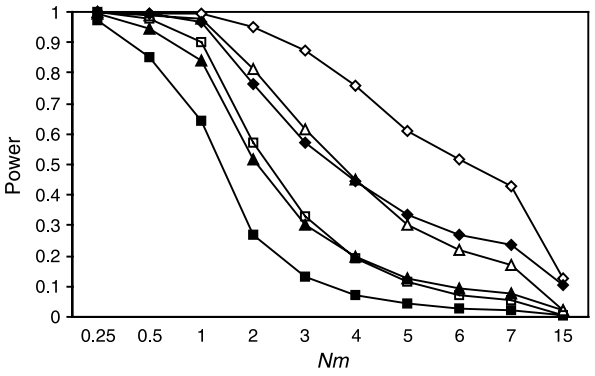

Fig. 3 A comparison of power when using L_h (filled symbols) or Λ (open symbols) as the test statistic, with α at 0.05 (diamonds), 0.01 (triangles) or 0.002 (squares). Simulated sample files as in Fig. 1.

It was not surprising to find that using L_h as a test statistic resulted in lower power than was observed with Λ (Fig. 3). However, the more interesting issue is the magnitude of this loss of power, and what the practical implications of that drop are. A useful way to summarize the difference between L_h and Λ, as observed in the case we examined, is that very similar power could be achieved with the two statistics by setting α, and thus type I error, five times higher for L_h than for Λ. For instance, we have found that, using Λ as test statistics, more than 90% of F_0 immigrants could be identified (on average) for $Nm = 1$ and $\alpha = 0.002$. When using L_h as the test statistic, this level of power could only be achieved under the permissive conditions of $\alpha = 0.05$, at which point we expect 2.45 type I errors (49 residents $*$ 0.05) per population, considerably outnumbering the single F_0 immigrant in each population. It bears mention that, when using L_h, observed type I error exceeded expectation at high values of Nm (not shown). However, if the estimate of type I error had been conservatively based on the entire sample ($50 * \alpha$), as will be the case when the actual number of immigrants and residents is unknown, then error rates would have been very close to expectation.

The effect of the number of sampled individuals on power was no surprise, with power increasing from a low of 0.32 at $s = 12$ to a maximum of 0.42 when s was 60 or higher (Fig. 4). Power did not rise as s increased beyond 60, although type I error decreased. The more surprising result was that the observed rate of type I error exceeded expectation when s was less than 60, peaking at more than double the expected rate when s was 12. In their simulation study based on a closed population model (i.e. without migration), Cornuet *et al*. (1999) experienced a similar bias

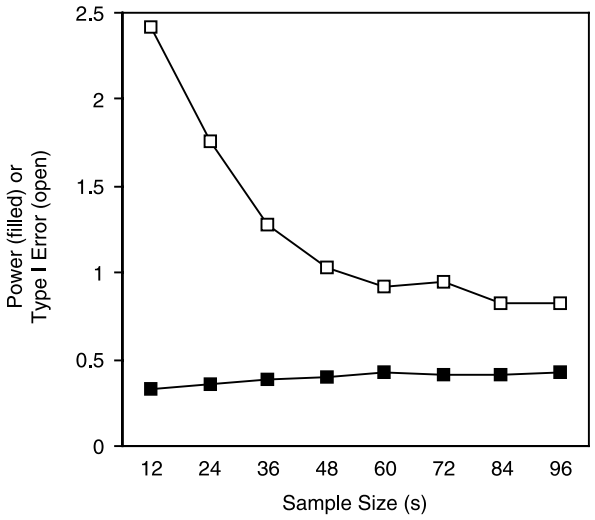

Fig. 4 The effect of sample size on power (filled symbols) and type I error rate (open symbols) relative to expectation [$\alpha * N * (1\text{-}m)$]. $N = 96$, $\mu = 0.005$, $l = 10$, $\alpha = 0.01$.

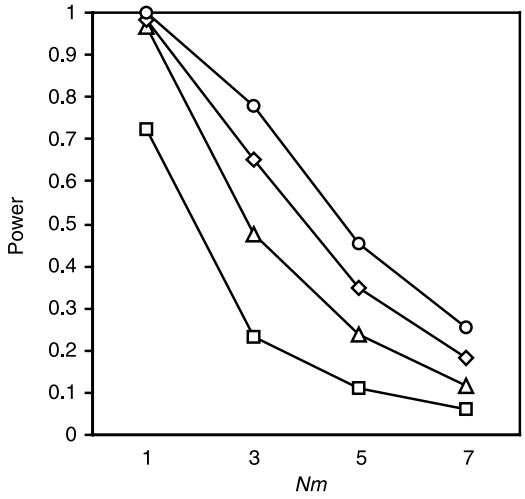

Fig. 5 Power observed with 5 (squares), 10 (triangles), 15 (diamonds) or 20 (circles) loci. $N = s = 50$, $\mu = 0.0025$, $\alpha = 0.01$.

associated with small sample sizes. Similarly, Rannala & Mountain (1997) gave an empirical example in which three of 12 aboriginal Australians (i.e. $s = 12$) were identified as having recent immigrant ancestry. When we reanalysed this file setting their program (IMMANC) to consider only the current generation (as we are doing in this paper), all three were identified as F_0 immigrants ($\alpha = 0.05$). This biologically surprising result could be caused in part by a small sample size bias, in addition to an excess of Type I errors associated with Monte Carlo resampling methods (Fig. 1). The data in Fig. 4, and the results of other explorations (data not shown), suggest that sample size will need to be close to 50 diploid individuals to avoid this bias for large populations.

The last sampling parameter that we explored was the number of loci sampled (l). With five loci, it may not be possible to achieve the combination of power and immigration rate required to make this technique practical. However, we observed a dramatic increase in power by doubling the number of loci, and a similar increase by doubling the number again (to 20; Fig. 5). This is very similar to results obtained in empirical assignment test studies of polar bears, where the rate of correct assignment was 10% higher with 16 loci than eight loci, and 10% higher with eight loci than with four loci (Paetkau *et al*. 1999). In practice, one would be able to adjust the number of loci to suit research needs and circumstances. For example, we observed a nearly 50% decrease in power when going from $Nm = 3$ to $Nm = 5$ with 10 loci, but this decrease could be almost completely offset by using 20 loci at $Nm = 5$.

We have been reporting power as the proportion of F_0 immigrants falling past the critical value. If one wishes to estimate migration rate by simply counting up the number of individuals falling past the critical value (minus the number of expected type I error), this is the most relevant presentation of data. However, another type of question would be to ask if there are biological differences, such as age, size or sex ratio, between F_0 immigrants and residents. To answer such questions, it may be sufficient to identify two groups of individuals, one with a preponderance of F_0 immigrants and the other comprising primarily residents. In creating such groups, there may be a need to maximize the proportion of immigrants in the group falling past the critical value, at the expense of power. In statistical terms, this is equivalent to minimizing the false discovery rate. Some statistical procedures have been proposed to do this in a reproducible manner (e.g. Benjamini & Hochberg 1995; Sabatti *et al*. 2003). Our simulations indicate that more stringent critical values were best for this purpose. For example, in Fig. 2 ($\alpha = 0.002$ with missing alleles at 0.01) we witnessed a rapid decrease in power as Nm increased beyond 1, but the proportion of F_0 immigrants in the group that fell past the critical value remained in excess of 0.9 up to $Nm = 5$ because there were so few type I errors. In fact, while the figure would suggest little hope for the general method at $Nm = 15$, 60% of individuals falling past the critical value at this point were still F_0 immigrants. Thus, even in extreme situations, it should be possible to identify two groups, one made up of a majority of F_0 immigrants and the other of a majority of residents. Any of the data in our figures can be expressed in these terms by calculating the expected number of type I errors [$\alpha * N * (1 - m)$] and the number of F_0 immigrants successfully identified (power $*$ Nm).

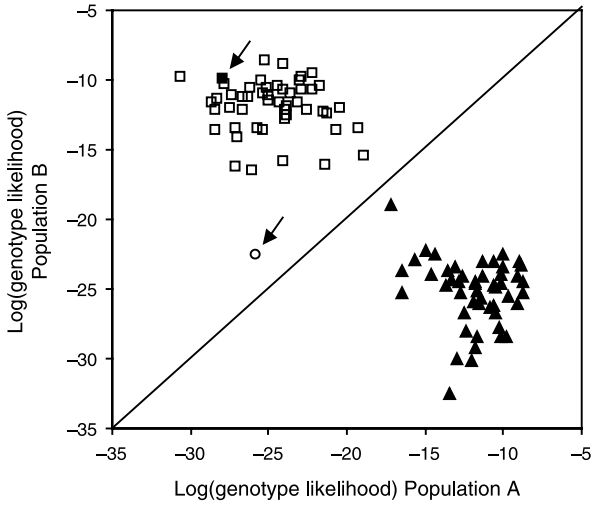

Fig. 6 Assignment plot for a pair of populations where the power to identify *F0* immigrants is high ($Nm = 1$, power = 0.984). The plots show individuals 'born' in three of the 10 populations in the system (squares, diamonds and circles, respectively), and 'captured' in two of those populations (shaded and open symbols, respectively). *F0* immigrants are indicated with arrows. The value on the *X*-axis is L_h for population A, and that on the *Y*-axis is L_h for population B. D_{LR} is the mean distance of individuals from the diagonal centre line, which distance could be used as a test statistic for a two population scenario (*a la* Rannala & Mountain 1997). $N = s = 50$, $\mu = 0.0025$, $l = 15$.

Predicting power

A researcher beginning the analysis of a data set is unlikely to know *N*, *Nm* or any of the other parameters that might affect power. Therefore, it would be helpful to have a way to estimate how much power one is likely to have with a given system of samples and markers. A good starting point for predicting power is to plot genotype likelihoods for the individuals from a given pair of populations (Fig. 6). If there is a clear separation between the two clusters of individuals, then one will feel confident in one's ability to identify animals born in one of these populations and captured in the other, and there may be no need for any more sophisticated analysis. However, in the more likely event that the resolution is not perfect, it would be useful to have a quantitative method of predicting power. For example, Cornuet *et al*. (1999) suggested that F_{ST} might be useful in this regard.

To address this issue, we recorded four measures of population differentiation during our simulations (F_{ST}, D_S, $(\delta\mu)^2$ and D_{LR}), and their utility in different contexts was explored using data sets where $N = s = 50$, and where l varied from 5 to 20 and the mutation rate (μ) from 0.0025 to 0.02. Normally one would not want a genetic distance measure to be affected by l, and indeed we found that the mean values of F_{ST}, D_S and $(\delta\mu)^2$ were insensitive to changes in this parameter (not shown). By contrast, D_{LR} increased as l, and thus differences in likelihood, increased between pairs of populations. While this is an undesired quality in the context of measuring inherent genetic differentiation between populations, it may be an asset in the context of predicting the power of assignment methods.

In a population model without geneflow, the expectation of D_S and $(\delta\mu)^2$ is the product of 2μ and the time since isolation (Nei 1972; Goldstein *et al*. 1995). Thus, it is no surprise to see that these metrics produced larger values at higher μ. Similarly, since assignment methods are better able to resolve between populations at higher μ (Cornuet *et al*. 1999; Estoup *et al*. 1998), it was no surprise that D_{LR} increased with increases in μ. By contrast, F_{ST} showed almost no sensitivity to μ at $Nm = 7$, although at $Nm = 1$ mean F_{ST} was considerably lower at $\mu = 0.02$ ($F_{ST} = 0.105$) than at $\mu = 0.0025$ ($F_{ST} = 0.158$), with intermediate values at intermediate values of μ; a result consistent with previous theoretical studies (e.g. Weir & Cockerham 1984; Rousset 1996).

Another consideration when comparing measures of differentiation is precision. Across the series of data sets where we compared distance metrics (28 series of 100 data sets, each with 10 populations), the mean coefficient of variation was relatively high for $(\delta\mu)^2$ (0.604), intermediate for D_S (0.145) and lowest for F_{ST} and D_{LR} (0.099 and 0.095, respectively). In short, if one wished to have a metric that accurately reflected connectivity (*Nm*) between the populations we studied, there can be no doubt that F_{ST} would be the measure of choice. However, while F_{ST} tracked *Nm* relatively well, it did not reflect the differences in power that resulted from changing l or μ. Thus, while F_{ST} has considerable value as an estimator of *Nm*, it performs less well as a predictor of power to detect F_0 immigrants.

To assess the performance of distance metrics as predictors of power, we plotted power against genetic distance for the 28 data series under consideration (Fig. 7). This presentation made it clear that D_{LR} performed best at predicting power. In the data sets that we considered, a value of D_{LR} in excess of 5 was always associated with near maximum power to distinguish immigrants and residents, while a value below 3 was associated with power of less than 0.5. This pattern held when examining other data sets described in this paper. We suggest that a combination of calculating D_{LR}, which provides the mean log likelihood ratio for a pair of populations, and plotting genotype likelihoods, which gives a visual indication of variation around that mean, will provide useful preliminary information on whether there is enough power in a given system to make detection of *F0* immigrants with assignment methods feasible, and hence direct, real-time estimate of migration rates.

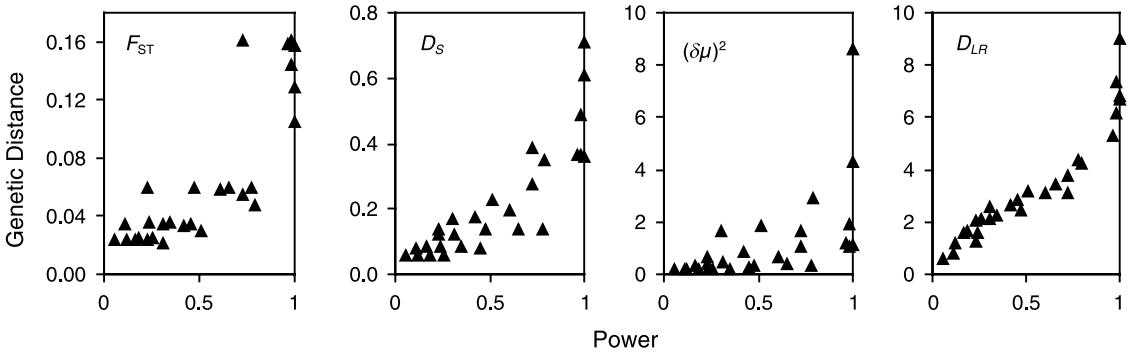

Fig. 7 Performance of four population genetic statistics in predicting power to identify *F0* immigrants. The 28 data sets involved different combinations of l (5–20), μ (0.0025–0.02) and Nm (1, 3, 5 or 7).

Acknowledgements

We thank Sylvain Piry for program coding, Bruce Rannala for sharing programs and data, and Laurent Excoffier for useful discussions about forward and backward simulation methods. This collaborative work was initiated thanks to Craig Moritz who welcomed three of the four authors into his lab at the University of Brisbane and provided financial and logistical support. David Paetkau was supported by the Natural Sciences and Engineering Research Council of Canada, Arnaud Estoup was supported by a grant from the Institut National de Recherche Agronomique devoted to the development of methods for real-time population genetics, and Michael Burden was supported by the WebPop consortium.

References

Banks MA, Eichert W, Olsen JB (2003) Which genetic loci have greater population assignment power? *Bioinformatics Application Note*, **11**, 1436–1438.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series*, **57**, 289–300.

Cornuet J-M, Piry S, Luikart G, Estoup A, Solignac M (1999) New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics*, **153**, 1989–2000.

Davies N, Villablanca FX, Roderick GK (1999) Determining the source of individuals, multilocus genotyping in nonequilibrium population genetics. *Trends in Ecology and Evolution*, **14**, 17–21.

Efron B (1983) Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*, **78**, 316–331.

Eldridge MDB, Kinnear JE, Onus ML (2001) Source population of dispersing rock-wallabies (*Petrogale lateralis*) identified by assignment tests on multilocus genotypic data. *Molecular Ecology*, **10**, 2867–2876.

Estoup A, Angers B (1998) Microsatellites and minisatellites for molecular ecology: theoretical and empirical considerations. In: *Advances in Molecular Ecology* (ed. Carvalho G), pp. 55–86. IOS Press, Amsterdam, The Netherlands.

Estoup A, Rousset F, Michalakis Y, Cornuet J-M, Adriamanga M, Guyomard R (1998) Comparative analysis of microsatellite and allozyme markers: a case study investigating microgeographic differentiation in brown trout (*Salmo Trutta*). *Molecular Ecology*, **7**, 339–353.

Favre L, Balloux F, Goudet J, Perrin N (1997) Female-biased dispersal in the monogamous mammal *Crocidura russula*: evidence from field data and microsatellite patterns. *Proceedings of the Royal Society of London Series B*, **264**, 127–132.

Goldstein DB, Ruiz Linares A, Feldman MW, Cavalli-Sforza IL (1995) Genetic absolute dating based on microsatellites and the origin of modern humans. *Proceedings of the National Academic Science USA*, **92**, 6723–6727.

Guinand B, Topchy A, Page KS, Burnham-Curtis MK, Punch WF, Scribner KT (2002) Comparison of likelihood and machine learning methods of individual classification. *Journal of Heredity*, **93**, 260–269.

Hudson RR (1991) Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, **7**, 1–44.

Leblois R, Estoup A, Rousset F (2003) Influence of mutational and sampling factors on the estimation of demographic parameters in a continuous population under isolation by distance. *Molecular Biology and Evolution*, **20**, 491–502.

Maudet C, Miller C, Bassano B *et al.* (2002) Microsatellite DNA and recent statistical methods in wildlife conservation management, applications in Alpine ibex [*Capra ibes (ibex)*]. *Molecular Ecology*, **11**, 421–436.

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

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

Ohta T, Kimura M (1973) A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetical Research*, **22**, 201–204.

Paetkau D, Amstrup SC, Born EW *et al.* (1999) Genetic structure of the world's polar bear populations. *Molecular Ecology*, **8**, 1571–1584.

Paetkau D, Calvert W, Stirling I, Strobeck C (1995) Microsatellite analysis of population structure in Canadian polar bears. *Molecular Ecology*, **4**, 347–354.

Paetkau D, Shields GF, Strobeck C (1998) Gene flow between insular, coastal and interior populations of brown bears in Alaska. *Molecular Ecology*, **7**, 1283–1292.

Paetkau D, Waits LP, Clarkson PL, Craighead L, Strobeck C (1997) An empirical evaluation of genetic distance statistics using microsatellite data from bear (Ursidae) populations. *Genetics*, **147**, 1943–1957.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Rannala B, Mountain JL (1997) Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences of the USA*, **94**, 9197–9201.

Rousset F (1996) Equilibrium values of measures of population subdivision for stepwise mutation process. *Genetics*, **142**, 1357–1362.

Sabatti C, Service S, Freimer N (2003) False discovery rate in linkage and association genome screens for complex disorders. *Genetics*, **163**, 829–833.

Stephens JC, Briscoe D, O'Brien SJ (1994) Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. *American Journal of Human Genetics*, **55**, 809–824.

Waser PM, Strobeck C (1998) Genetic signatures of interpopulation dispersal. *Trends in Ecololgy and Evolution*, **13**, 43–44.

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

Whitlock MC, McCauley DE (1999) Indirect measures of gene flow and migration, $F_{ST} \neq 1/(4Nm + 1)$. *Heredity*, **82**, 117–125.

Wiens JA (1997) Metapopulation dynamics and landscape ecology. In: *Metapopulation Biology* (eds Hanski IA, Gilpin MW), pp. 43–63. Academic Press, San Diego.

Wilson GA, Rannala B (2003) Bayesian inference of recent migration rates using multilocus genotypes. *Genetics*, **163**, 1177–1191.

Woods JG, Paetkau D, Lewis D, McLellan BN, Proctor M, Strobeck C (1999) Genetic tagging of free-ranging black and brown bears. *Wildlife Society Bulletin*, **27**, 616–627.

Wright S (1931) Evolution in Mendelian populations. *Genetics*, **16**, 97–159.