Molecular Ecology (2003) **12**, 1979–1991 doi: 10.1046/j.1365-294X.2003.01856.x

AFLP utility for population assignment studies: analytical investigation and empirical comparison with microsatellites

DAVID CAMPBELL, PIERRE DUCHESNE and LOUIS BERNATCHEZ
Département de Biologie, Université Laval, Ste-Foy, Québec, Canada, G1K 7P4

Abstract

Individual-based population assignment tests have thus far mainly relied on the use of microsatellite loci. However, the logistic difficulty of screening large numbers of loci required to reach sufficient statistical power hampers the usefulness of microsatellites in situations of weak population structuring. Amplified fragment length polymorphisms (AFLP) represents an alternative for overcoming this logistical issue as the technique allows the user to characterize a much larger number of loci with a comparable analytical effort. In this study, an assignment test based on maximum likelihood for dominant markers was used to investigate the potential usefulness of AFLP for population assignment. We also compared assignment success achieved with AFLP with that obtained using microsatellites in a case study of low population differentiation involving whitefish (*Coregonus clupeaformis*) sympatric ecotypes. The analytical investigation showed that the minimum number of AFLP loci required to reach an assignment success of 95% stood within values that are easily achievable in many situations. This also showed how assignment success varied according to the number of AFLP loci used, their absolute frequency and their frequency differential and sampling errors, as well as the number of putative source populations. The case study showed that given a comparable analytical effort in the laboratory, AFLP were much more efficient than the microsatellite loci in discriminating the source of an individual among putative populations. AFLP resulted in higher assignment success at all levels of stringency and the log-likelihood differences between populations obtained with AFLP for each individual were much larger than those obtained with microsatellites. These results indicate that research involving individual-based population assignment methods should benefit importantly from the use of AFLP markers, especially in systems characterized by weak population structuring.

Keywords: AFLP, assignment test, *Coregonus clupeaformis*, likelihood, microsatellites, population assignment

Received 5 December 202; revision received 26 February 2003; accepted 26 February 2003

Introduction

The individual-based population assignment test was introduced first by Paetkau *et al*. (1995) as a means to quantify degrees of genetic differentiation among populations. Subsequently, this method has proved useful in a wide variety of applications in population and conservation biology (reviewed in Waser & Strobeck 1998; Davies *et al*. 1999; Hansen *et al*. 2001). Assignment tests may thus be used in both plants and animals to establish relationships among individuals within and among populations or higher taxonomic groupings (Estoup *et al*. 1995; Cornuet *et al*. 1996; Ellegren *et al*. 1996; Estoup & Anger 1998; Roques *et al*. 1999) and to identify putative source populations of individuals of unknown origin. This, in turn, has enabled researchers to determine the relative contribution of each potential source population in mixed fisheries (Roques *et al*. 1999; Potvin & Bernatchez 2001), to assign individuals during migration (Haig 1997; Mountain & Cavalli-Sforza 1997; Palsboll *et al*. 1997; Rannala & Mountain 1997; Eldridge *et al*. 2001), to estimate sex-biased dispersal (Favre *et al*. 1997; Dallimer *et al*. 2002) and gene

Correspondence: L. Bernatchez. Fax: 418 656 2043;
E-mail: Louis.Bernatchez@bio.ulaval.ca

flow (Paetkau *et al*. 1998), to identify potential admixture between populations (Nielsen *et al*. 1997) and to estimate the long-term effects of population stocking (Hansen 2002). Forensic sciences have also benefited from assignment tests as they can be used to discriminate if an animal originates from an illegal source (Primmer *et al*. 2000).

Thus far, assignment tests have relied mainly on the use of microsatellite loci. Although the resolution obtained with these markers is often adequate, an important constraint faced by the use of microsatellites in any type of individual-based population assignment is the lack of statistical power in situations of weak population differentiation ($F_{ST} < 0.05$) (Cornuet *et al*. 1999). In such situations, assignment success also decreases rapidly with increased stringency on assignment decision (Roques *et al*. 1999). As limited statistical power in situations of weak differentiation between putative source populations hampers our ability to trace back the population membership of a set of individuals, improvement of present-day methods is required in order to enhance actual resolution in population assignment studies (Hansen *et al*. 2001).

A main ingredient towards improving current population assignment methods is increasing the number of loci used in such studies (Bernatchez & Duchesne 2000). Herein, the major limitation inherent with microsatellite markers resides in the logistic difficulty of increasing the number of useful loci for assignment tests. Indeed, developing and applying large numbers of microsatellite markers may be technically challenging, expensive and time-consuming, particularly for species for which no previous marker development was undertaken (Goldstein & Pollock 1997). Amplified fragment length polymorphisms (AFLP) represents an alternative towards overcoming this logistical issue as the technique generates a large number of loci. Also, the cost and time required for this moderately challenging protocol is relatively low (Vos *et al*. 1995; Rieseberg 1998; Mueller & Wolfenbarger 1999). The large number of markers generated by AFLP can then be screened to select the best set of loci required in assignment procedures. On the other hand, the low allelic diversity exhibited by AFLP markers (di-allelic) may be viewed as a potential constraint towards increasing statistical power relative to that achieved with microsatellite markers, which commonly have more than 10 alleles per locus (Neff & Gross 2001). However, Bernatchez & Duchesne (2000) demonstrated analytically that in contrast to studies of parentage analysis, increasing the number of loci used is more critical than increasing allelic diversity per locus in studies of population assignment. Thus, assessing whether many (e.g. > 100) di-allelic AFLP loci would be more efficient than using few (e.g. < 10) hypervariable microsatellite loci for population assignment would offer critical insights towards the improvement of assignment procedures.

In this context, the general objective of this paper was to use the maximum likelihood method of Paetkau *et al*. (1995), adapted for use with dominant markers in order to investigate the usefulness of AFLP loci for population assignment, and compare empirically assignment success achieved with both AFLP and microsatellite markers in a case study. We describe first the assignment method based on AFLP genotype likelihoods. The main focus is on the minimum number of loci to reach predefined levels of probability of correct assignment. Using this method, we investigate the magnitude of effect that sampling errors would have on assignment success. We then explore the effect of the number of candidate populations on the minimum number of loci required to reach predefined assignment success levels.

Secondly, we present empirical results from a pair of lake whitefish (*Coregonus clupeaformis*) sympatric ecotypes that represent a case of low population genetic differentiation. Using six microsatellite loci, Lu & Bernatchez (1999) found a negative correlation between the extent of gene flow and morphological specialization between lake whitefish ecotypes in different lakes, with observed F_{ST} estimates as high as 0.25 in some lakes. In East Lake (Southeastern Québec, Canada), dwarf and normal ecotypes are differentiated morphologically despite the persistence of a high gene flow between them ($F_{ST} = 0.058$, SD = 0.013–0.110; $Nm = 4.06$, SD = 2.02–18.98). Given the previous fine-scale genetic data analyses available for the species, the East Lake dwarf and normal whitefish pair represents an appropriate template to explore the potential of AFLP for population assignment relative to that of microsatellites. In this perspective, assignment tests were used to explore the effect of: (i) selective primer combination and (ii) number of loci. The assignment success obtained with AFLP markers under various stringency levels were then compared with those obtained from microsatellite markers in order to contrast their discriminating power.

Analytical investigation

Assignment procedure

Formally, AFLP genotypes are strings of presences (1 s) and absences (0 s) of fragments. Distinct populations are expected to show differences in their respective frequencies of presence among polymorphic loci. These differences can be used to assign specimens to their original population based on the individual's AFLP genotype, given a set of candidate populations.

Let us call G the genotype of the individual to be assigned. We first compute the likelihood that G be found in each of the candidate populations based on their respective presence frequencies. G is then assigned to the population showing the highest likelihood for G (Paetkau *et al*.

1995). The likelihood of genotype G in population X, among all possible genotypes in *X*, is the product:

$$L_X = \prod_{i \in S_P} f_{i,X} \prod_{j \in S_A} (1 - f_{j,X}) \quad (1)$$

where

$f_{i,X}, f_{j,X}$, = frequency of presence in locus *i*, *j* in population X
S_P = set of indices of loci with presences in genotype G
S_A = set of indices of loci with absences in genotype G

For computational reasons, it is more convenient to express this likelihood relationship in log terms, such that:

$$\log L_X = \sum_{i \in S_P} \log (f_{i,X}) + \sum_{j \in S_A} \log (1 - f_{j,X}) \quad (2)$$

Practically, this means that each presence (1) in G will be replaced by the log of the frequency of 1 s within the same locus in population X and each absence (0) by the log of the frequency of 0 s. Because log(0) is not defined, it is necessary to replace frequencies of zero by some appropriate value. We found that $1/(n + 2)$ was the most appropriate substitution value, where *n* is the sample size (see Appendix I). Subsequently, log values are summed over loci to obtain the log-likelihood of genotype G being assigned to X. Log-likelihoods are computed for each population. In order to make the decision of either assigning genotype G to one population or not assigning it at all, different stringency levels can be applied. The stringency level is defined as the absolute minimal difference between the largest and next to largest log-likelihood of genotype G that is required to assign the individual to one among the putative populations ($|\log L_A - \log L_B| > \varepsilon$, where $\varepsilon \geq 0$). Four stringency levels are used commonly ($\varepsilon = 0, \varepsilon = 1, \varepsilon = 2$ and $\varepsilon = 3$). The $\varepsilon = 1$, $\varepsilon = 2$ and $\varepsilon = 3$ levels, respectively, mean that a multilocus genotype has to be 10, 100 or 1000 times more likely in one population than in any other to be assigned. The $\varepsilon = 0$ level requires only that the genotype be more likely in one population relative to the other(s). If the log-likelihood difference is not equal to or higher than the selected stringency level, the genotype is not assigned and the procedure is said to have failed.

This method is an adaptation for dominant markers of Paetkau *et al.*'s method (1995) for codominant markers, which was also applied recently by Congiu *et al.* (2001). For microsatellites, each single-locus likelihood value is obtained by multiplication of the frequencies of its two component alleles. In the context of AFLP (dominant) markers, the only two possible one-locus-genotypes are 0 and 1. Consequently, their likelihoods are simply their respective frequencies within the locus. Equations 1 and 2 are based on the assumptions that $f_{i,X}$-values are accurate, and that the loci are statistically independent (no linkage disequilibrium).

Two-populations model

Assignment success probability. Loci with larger frequency differences should have more assignment power. Given that f_A and f_B are the respective frequencies of presence for populations A and B, the frequency differential ($FD = |f_A - f_B|$) is a natural candidate for measuring locus quality (Shriver *et al.* 1997). Given that numerous pairs (f_A, f_B) can produce identical frequency differential values, it is also of interest to assess whether such pairs have the same assignment power. For example, is (0.01, 0.15) a better or worse pair than (0.16, 0.30)? In order to address this issue, we used a simple frequency distribution model in which f_A and f_B were held constant across all loci. For the sake of simplicity but without any loss of generality we also assumed f_B larger than f_A. This model allowed us to derive a straightforward procedure (see Appendix I) to compute exact probabilities of correct assignment, given any number of loci (*l*) and any constant frequency pair (f_A, f_B).

We first assessed how the probability of correct assignment was affected by the frequency f_A, the frequency differential, and the number of loci under the restrictions of the simple model. The probability of correct assignment was computed as a function of the frequency differential (*FD*) ranging from 0.01 to 0.250, for each of f_A (0.001, 0.010, 0.020, 0.050, 0.100, 0.150, 0.200, 0.250) and each number of loci (20, 50, 100) for the perfect situation in which there is no sampling error and no error in assigning a genotype to an individual (no scoring error or polymerase chain reaction (PCR)-induced artefacts). Note that the computation of probabilities of correct assignment within the two-populations model was not based on simulations but obtained analytically.

The probability of correct assignment increased with the frequency differential (*FD*) and the number of loci (Fig. 1). The probability of correct assignment also increased sharply as f_A approached zero. The same effect has been observed with f_B close to 1 (data not shown). These results demonstrate that the quality of loci depends not only on the frequency differential of the pair (f_A, f_B), but also on the proximity of f_A or f_B to either one of the tail ends of the (0, 1) probability interval. In order to investigate the effect of the number of loci on assignment power, we therefore retained only tail frequency pairs (0.001, f_B) and centre pairs (0.25, f_B). These two types, referred to below as *T* and *C* loci, provide lower and upper bounds for the minimum number of loci required to reach predefined levels of correct assignment.

Minimal number of loci. The minimum numbers of loci required to reach 90%, 95% and 99% probability of correct assignment (M_{90}, M_{95} and M_{99}) were computed for frequency differentials ranging from 0.05 to 0.70 for both T and *C* type loci. Minimal number of loci as a function of the

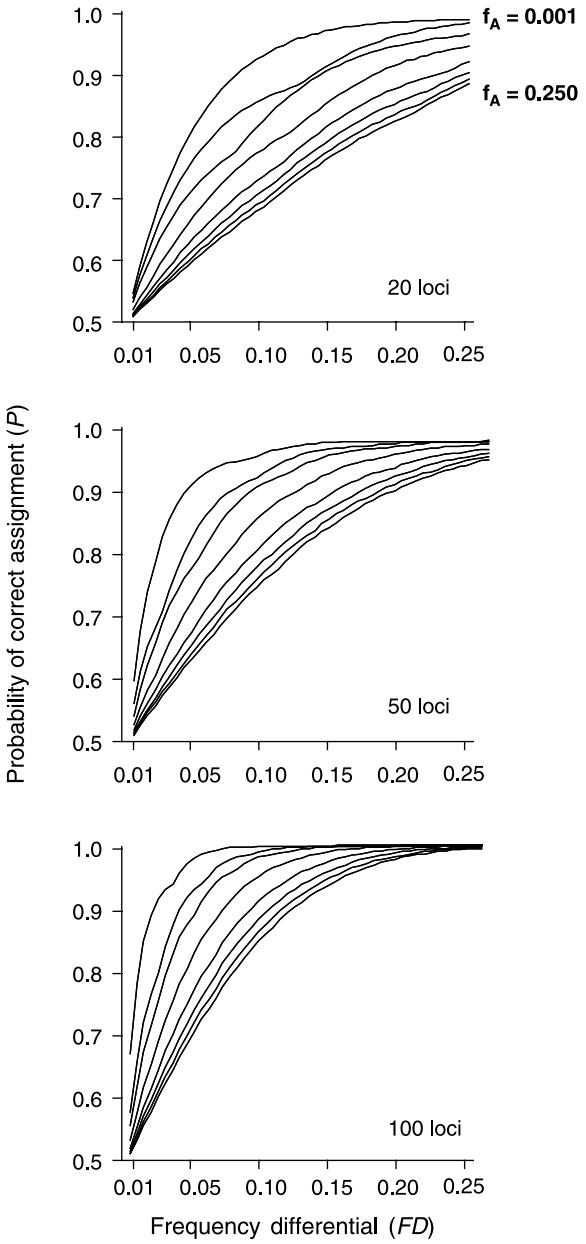

Fig. 1 Probability of correct assignment as a function of frequency differential ($FD = |f_B - f_A|$) for three different numbers of loci (20, 50, 100) with $\varepsilon = 0$. Curves are labelled according to f_A (0.001, 0.01, 0.02, 0.05, 0.10, 0.15, 0.20, 0.25) and lower curves correspond to higher values of f_A.

frequency differential revealed that the number of loci required rose sharply as the frequency differential decreased (Fig. 2). Curves obtained with *T*-loci show lower values of minimum number of loci but differences between the *T*- and *C*-type loci decreased as the frequency differential increased. For any probability of correct assignment, the minimum number of loci never exceeded 40 for frequency differentials equal or higher to 0.35 in *C*-loci and 0.15 in *T*-loci. Loci with lower frequency

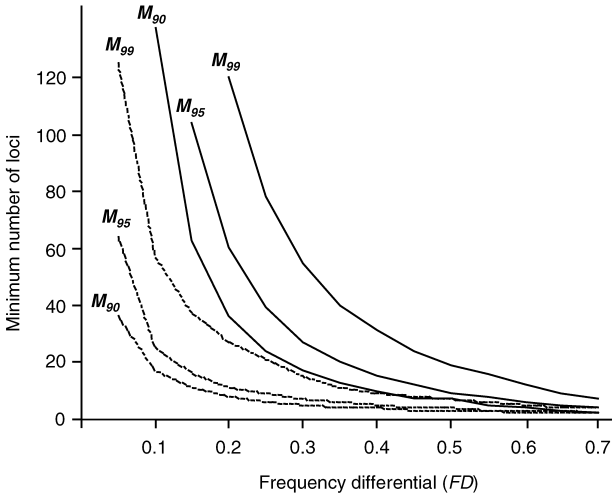

Fig. 2 Minimum number of loci as a function of frequency differential to reach a specific probability of correct assignment (M_{90}, M_{95} and M_{99}) with $\varepsilon = 0$. Dashed curves were obtained using *T*-loci ($f_A = 0.001$), whereas solid curves were obtained using *C*-loci ($f_A = 0.25$).

differential values, although less informative, can also provide substantial allocation power. Indeed, sets of approximately 60 *C*-loci with $FD = 0.2$ or sets of 60 *T*-loci with *FD* as small as 0.05 sufficed to reach an expected probability level of correct assignment of 95%.

Effects of sampling errors on the probability of correct assignment. In the above two-populations model, we assumed no sampling error on frequency estimates, i.e. $\phi_A = f_A$ and $\phi_B = f_B$. As the latter assumption is unrealistic and errors on presence frequency estimates will certainly have a negative impact on the probability of correct assignments (*P*) (Smouse & Chevillon 1998), it is essential that we assess the effect of this loss of power on M_{90} for various sample sizes. We modelled the error on f, the population (real) presence frequency, under the assumption that, prior to the estimate ϕ, no relevant information is available on the parameter f. We also assumed that the sample size is small compared with the total population size so that the number of sampled presences is expected to be distributed binomially. Based on the previous assumptions, we were able to build the density function for f, given the sample size *n* and the estimate ϕ. To compare the relative power of our procedure with and without sampling error we generalized the above two-populations model. As before, the frequency estimates ϕ_A, ϕ_B were held constant across loci but the population frequencies f_A, f_B were obtained by sampling from the f-densities at each locus. Hence, each population was represented by an array **f** of sampled f_s. To obtain estimates of the expected proportion *P* of successful assignments as a function of estimates ϕ_A, ϕ_B and sample size *n*, 100 pairs (A, B) of populations were

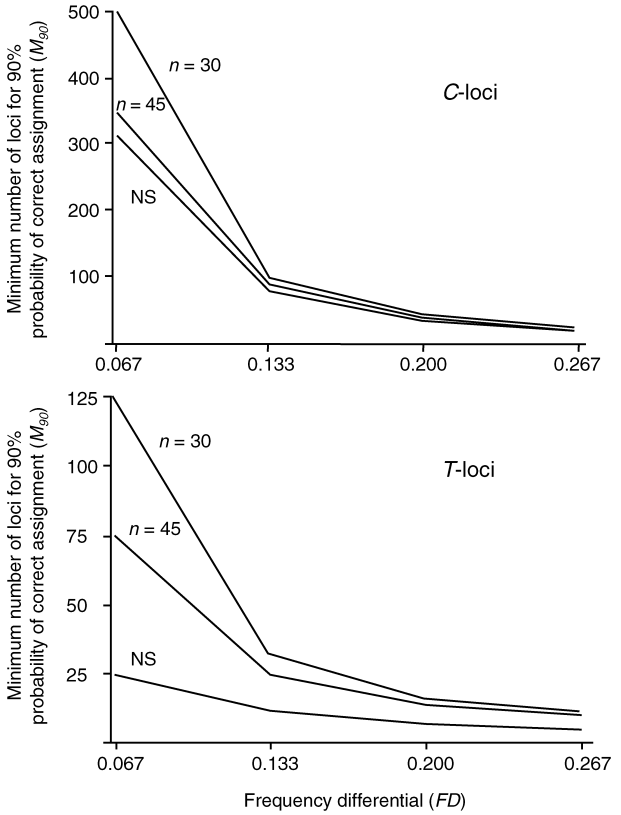

Fig. 3 Curves of minimum numbers of loci needed to reach 90% probability of correct assignment (M_{90}) as a function of frequency differential for the three error conditions ($n = 30$, $n = 45$, NS (no sampling error)). Top panel shows *C*-loci and bottom panel shows *T*-loci.

generated randomly and 1000 genotypes were generated randomly from each population. The details of the construction of the f-densities, the sampling of f and the computation of the expectancy of *P* are provided in Appendix II.

We calculated estimates of M_{90} for sample sizes $n = 30$ and $n = 45$. Pairs of (ϕ_A, ϕ_B) estimates were built from $\phi_A = 0$ (*T*-loci) and $\phi_A = 0.266$ (*C*-loci) and $FD = 0.067, 0.133, 0.200, 0.267$. Values of M_{90} without sampling error were computed analytically for the same set of (ϕ_A, ϕ_B). Figure 3 contains two panels (*T*- and *C*-loci) of M_{90} curves as a function of *FD* for $n = 30, 45$ and no sampling error (NS). As expected, the minimum number of loci needed to reach 90% probability of correct assignment was smallest with no sampling error and largest with the smaller sample size ($n = 30$) for both *T*- and *C*-loci. The most significant discrepancies in number of loci (M_{90}) between $n = 30$, 45 and no sampling error were found with the smallest frequency differential ($FD = 0.067$), but they decreased very rapidly with increasing frequency differentials (*FD*). Between the three error conditions, the *T*-loci showed more relative discrepancies in M_{90} than *C*-loci. With $FD \geq 0.133$, the relative discrepancies almost vanished in *C*-loci and absolute discrepancies never exceeded 20 in *T*-loci.

Although loci with $FD = 0.067$ do provide enough information to reach the 90% level of correct asssignments, they are much more sensitive to sampling error and so increasing *n* from 30 to 45 does have a dramatic effect on their assignment power. Indeed, with $n = 30$, 125 *T*-loci and 500 *C*-loci are needed to reach the targeted 90% correct assignment, whereas with $n = 45$ only 75 *T*-loci and 350 *C*-loci are needed. Therefore, loci with low frequency differentials, although more sensitive to sampling error, do provide significant assignment power when their presence frequencies are sampled adequately and their number is sufficiently high. However, with $n = 10$ and $FD = 0.1$, we failed to reach the 90% level of correct assignment even with very large numbers of loci. In fact, there were clear signs that E(*P*) reached an upper limit both with *T* and *C*-loci. These limits were approximately 0.86 (*T*-loci) and 0.76 (*C*-loci) and were reached with numbers of loci in the 100–150 range (graphs not shown).

Multiple populations model

We subsequently considered situations with more than two candidate populations and random frequency variation among the populations. We kept the notion of a fixed frequency differential ($FD = f_{max} - f_{min}$) where f_{max} and f_{min} represent the maximum and minimum frequencies among all populations, respectively. We also held f_{max} and f_{min} constant across all loci. Random frequency variation among populations was generated by the use of a Monte Carlo simulator. For each locus, $N-2$ frequencies, where *N* equals the number of populations, were generated randomly according to $U(f_{min}, f_{max})$, i.e. the uniform distribution with lower and upper bounds, respectively, equal to f_{min}, f_{max}. The two remaining frequencies were precisely f_{min}, f_{max}. The frequencies were then attributed randomly to the *N* populations. At this point, each population was defined by an array of *l* frequencies, where *l* equals the number of loci. Random genotypes were generated for each of the *N* populations by performing 1000 Bernoulli trials for each locus, thus generating 1000 genotypes per population. The number of iterations, each with a renewed set of *N* populations, was 100 for each pair of frequency bounds (f_{min}, f_{max}).

Minimal number of loci with more than two candidate populations. Using the Monte Carlo simulator described above, we sought the minimum number of loci to reach a probability of correct assignment of at least 90% (M_{90}). In the search for M_{90}, we considered *C*-loci (f_{min}, f_{max}) = (0.25, 0.45) (0.25, 0.55) (0.25, 0.65) (0.25, 0.75) (0.25, 0.85) (0.25, 0.95) for each of the following number of populations ($N = 3, 4, 5, 7, 10$). The random genotypes generated were

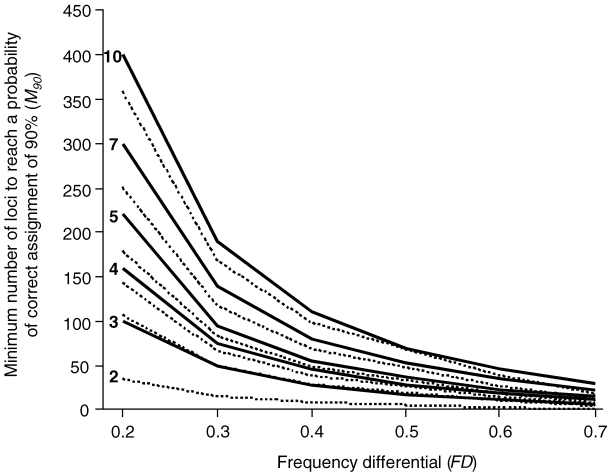

Fig. 4 Curves of minimum numbers of loci needed to reach 90% probability of correct assignment (M_{90}) as a function of frequency differential for multiple populations with *C*-loci ($f_A = 0.25$). Solid curves indicate simulation results (curves are labelled according to number of populations (N) = 3, 4, 5, 7, 10. Dashed curves represent estimates based on the relationship: $M_{90} = M_{90}$ for two populations*N (estimates were performed with N = 3, 4, 5, 7, 10). The dashed curve labelled with 2 represents M_{90} for two populations.

assigned to the N populations and the proportion of correct assignments was computed. The results obtained using the simulations for two populations, as a special case of the multiple populations model, were similar to those obtained from the two-populations (analytical) model in terms of the minimal number of loci needed to reach a probability of correct assignment of 90%, thus validating the calculation procedures of the multiple populations model.

With more than two populations, we found that M_{90} stood between N and $N + 1$ times the minimum number of loci for the two populations case (Fig. 4). Hence, for $N \geq 3$, a simple quasi-linear relationship appeared to hold between number of populations and minimum number of loci to reach a probability of correct assignment of at least 90% (M_{90}). For example, with *C*-loci and a frequency differential of 0.20, m^{90} for two populations is equal to 37 loci, and so M_{90} for 10 populations should fall between 370 ($37 \times N$) and 407 ($37 \times N + 1$) loci, which is indeed consistent with the simulated curve (400 loci, 10 populations, $FD = 0.20$, Fig. 4). However, as this relationship is based solely on observations from simulations, it should be considered cautiously for f_{min} values different from 0.25 and for probabilities of correct assignment different from 90%. The number of loci needed for reliable assignment with multiple populations appear relatively high. However, such numbers of loci (e.g. between 300 and 2000) have been screened commonly in previous studies using AFLP (e.g. Albertson *et al*. 1999; Rogers *et al*. 2001; Wilding *et al*. 2001; Young *et al*. 2001; Bensch *et al*. 2002; Orgen & Thorpe 2002).

Case study

Methodological outlines. AFLP analysis was performed on dwarf (n = 29) and normal (n = 29) lake whitefish sampled from East Lake, which is located within the secondary contact zone of the St-John river basin of northeastern North America (see Lu & Bernatchez 1999 for location map). DNA samples for AFLP analysis were the same as those used by Lu & Bernatchez (1999) for microsatellite analysis.

The AFLP plant mapping kit (Applied Biosystems, Inc.) was used according to the protocol of Vos *et al*. (1995) to generate AFLP markers. Three selective amplification primer combinations (*Eco*RIACA/*Mse*ICTA, *Eco*RIAGG/*Mse*ICTG and *Eco*RIACC/*Mse*ICTC) were used. For the remainder of the text, we will refer to these primer combinations using the *Eco*RIA (xx) and *Mse*IC (xx) selective nucleotide extension (e.g. *Eco*RIACA/*Mse*ICTA = CATA). The procedure for AFLP analysis has been detailed elsewhere (Rogers *et al*. 2001).

A descriptive analysis of our AFLP data set was conducted in order to characterize the loci generated by each primer combination. First, a distribution of frequency differential values was determined for each primer combination without any regard to tail proximity. Then, each locus was classified into one of three classes defined on the basis of tail proximity and frequency differential. Within the context of the case study, the loci will be characterized as *C*-loci when both presence frequencies f_A and f_B belong to the *centre* interval (0.01, 0.99) and otherwise as *T*-loci. *C*-loci with $FD < 0.20$ and *T*-loci with $FD < 0.05$ are categorized as the 'least' informative class, *C*-loci with $0.20 \leq FD < 0.35$ or *T*-loci with $0.05 \leq FD < 0.15$ figure as the 'good' class and *C*-loci with $FD \geq 0.35$ and *T*-loci with $FD \geq 0.15$ figure as the 'excellent' class. Given the analytical results, this classification gave prior indications as to the quality of each primer combination with regard to number of loci, frequency differential and tail proximity.

Following the method described above, reassignment tests were performed for each of the three AFLP primer combinations separately, as well as for the data set as a whole. In each case, the first test was conducted using all loci having a frequency differential higher than zero (minimal $FD = 0$). A 5% increment in the minimal frequency differential was added between each subsequent test to select the set of loci to be used. Tests were performed until no loci remained. In so doing, the number of loci used for reassignment decreased whereas the average quality of loci increased from the first to the last test. Each individual to be reassigned was removed from the source population when estimating the presence frequency of each locus. The purpose of this 'leave-one-out' procedure is to avoid an upward bias in assignment success (Smouse *et al*. 1982). Assignment success was expressed as the ratio of correct

Table 1 Classification of AFLP loci as a function of the frequency differential and tail proximity for each primer combination and the whole data set. The numbers of least informative, good and excellent markers are obtained based on the criteria defined in Methodological outlines. 'Least' informative loci = *T*-loci ($FD < 0.05$) + *C*-loci ($FD < 0.20$), 'good' loci = *T*-loci ($0.05 \leq FD < 0.15$) + *C*-loci ($0.20 \leq FD < 0.35$) and 'excellent' loci = *T*-loci ($FD \geq 0.15$) + *C*-loci ($FD \geq 0.35$).

	CATA	GGTG	CCTC	CATA and GGTG and CCTC
T-loci ($FD < 0.05$)	13	6	1	20
T-loci ($0.05 \leq FD < 0.15$)	25	11	16	52
T-loci ($FD \geq 0.15$)	7	5	0	12
C-loci ($FD < 0.20$)	35	3	11	49
C-loci ($0.20 \leq FD < 0.35$)	9	2	5	16
C-loci ($FD \geq 0.35$)	7	14	2	23
Number of least informative loci	48	9	12	69
Number of good loci	34	13	21	68
Number of excellent loci	14	19	2	35
Total number of loci	96	41	35	172
Average $FD \pm$ SD	0.14 ± 0.14	0.29 ± 0.25	0.12 ± 0.11	0.17 ± 0.18

assignments over all decisions, i.e. assignments plus nonassignments (failures).

All reassignment tests were conducted at four different levels of stringency ($\varepsilon = 0$, $\varepsilon = 1$, $\varepsilon = 2$ and $\varepsilon = 3$) in order to evaluate the discriminating power of different sets of AFLP loci. Increasing values of ε will lead generally to an increased proportion of failures and thereby to a decreased overall proportion of both correct and incorrect assignments. Therefore, the rationale in using different stringency levels to evaluate the discriminating power of a set of loci is that the higher the assignment success at high levels of stringency, the higher the discriminating power of a set of loci. All procedures with AFLP were carried out using AFLPOP (Duchesne & Bernatchez 2002), a macro-Excel based software available free of charge on the following website: http://www.bio.ulaval.ca/contenu-fra/professeurs/Prof-l-bernatchez.html.

Individuals used in this study were genotyped previously at six microsatellite loci (*Bwf1*, *Bwf2*, *C2-157*, *C4-157*, *Cocl23* and *Cocl22*) (methodological details in Lu & Bernatchez 1999), allowing us to compare AFLP with microsatellite-based assignments. Overall, comparable analytical effort in the laboratory was required to collect the data from both methods on whitefish populations. Thus, running one AFLP 30-lane gel on a 377 automated DNA sequencer (Applied Biosystems, Inc.) allowed genotyping of 30 individuals with a multiplex of three primer combinations, whereas the equivalent microsatellite gel allowed genotyping of 30 individuals at three microsatellite loci. Two PCR amplifications per individual analysed were necessary for both methods (preselective and selective amplification for AFLP and two multiplexes amplifications for microsatellites).

Results of genetic polymorphism and population differentiation using microsatellite loci are reported from Lu & Bernatchez (1999), where details on the analysis are also presented. Individuals were reassigned on the basis of their microsatellite multilocus genotype using the Bayesian method in GENECLASS (Cornuet *et al*. 1999). Assignment success was estimated at four levels of stringency ($\varepsilon = 0$, $\varepsilon = 1$, $\varepsilon = 2$ and $\varepsilon = 3$) in order to compare the discriminating power of AFLP loci and microsatellite loci.

Results and discussion

Effect of the primer combination on assignment success

The characterization of the AFLP data set is summarized in Table 1 and Fig. 5. A total of 172 of 182 AFLP loci were informative ($FD > 0$) with an average frequency differential of 0.17. Primer combinations CATA, GGTG and CCTC yielded 96, 41 and 35 informative loci, respectively. CATA generated the largest number of 'least' informative loci (48 for CATA, 12 for CCTC and nine for GGTG) and of 'good' loci (34 for CATA, 21 for CCTC and 13 for GGTG). GGTG generated the largest number of 'excellent' loci (19 for GGTG, 14 for CATA, two for CCTC) and had the highest average frequency differential ($FD = 0.29$ for GGTG, 0.14 for CATA and 0.12 for CCTC). Using all loci available for each primer combination, very high assignment success was obtained at $\varepsilon = 0$ with CATA (100%), GGTG (95%) or CCTC (97%). When increasing the stringency level to $\varepsilon = 3$, the assignment success decreased from 100% to 86% with CATA, from 95% to 84% with GGTG and from 97% to 29% with CCTC. When using the complete data set, increasing the stringency level to $\varepsilon = 3$ increased the rate of failures by only 3% relative to $\varepsilon = 0$, thus providing an assignment success of 97%.

Thus, contrary to what was expected from the 'random' nature of the amplification of AFLP loci (Vos *et al*. 1995), the three AFLP primer combinations used in this study provided sets of loci varying significantly in number and quality (frequency differential and tail proximity) of loci. As a result, variable assignment success was achieved with these three primer combinations with CATA providing the

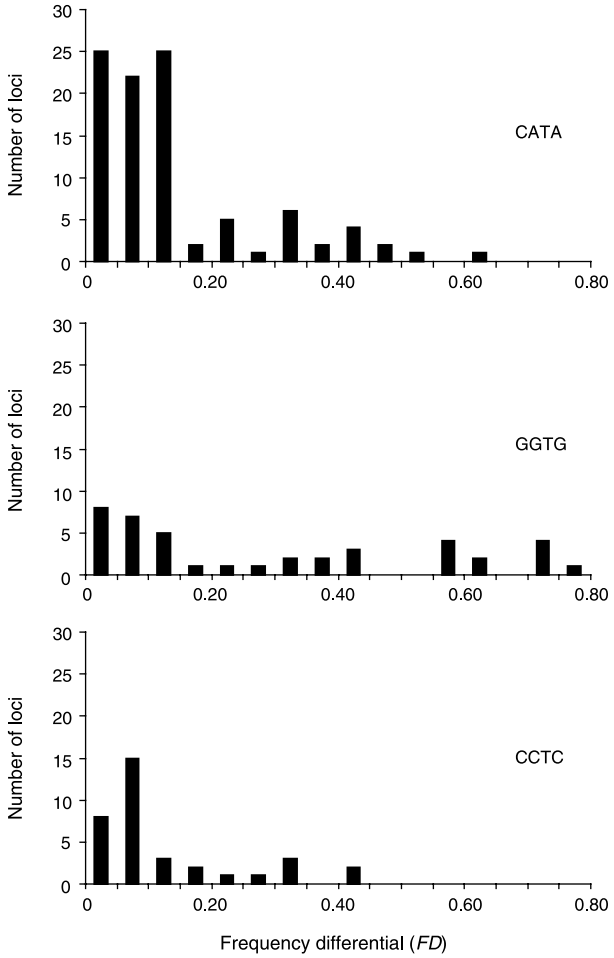

Fig. 5 Distribution of frequency differential values observed in CATA, GGTG and CCTC primer combinations.

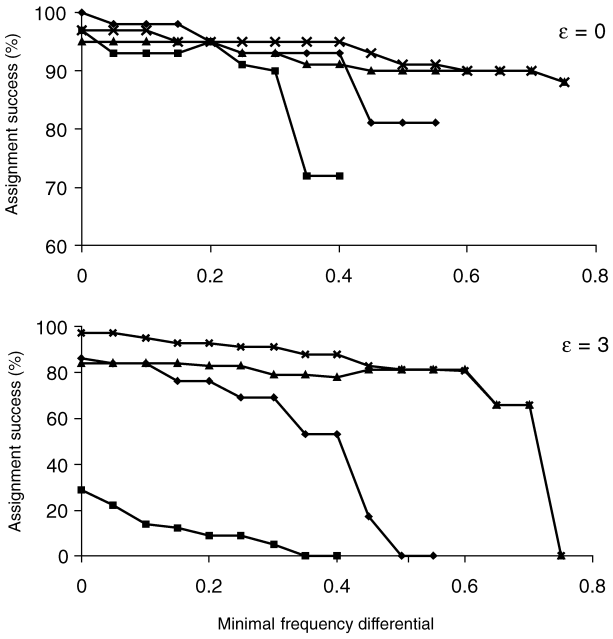

Fig. 6 Assignment success as a function of minimal frequency differential for two stringency levels (ε = 0 and ε = 3): (◆), CATA (▲) GGTG, (■) CCTC, (×) three primer combinations combined.

strongest and CCTC the weakest discriminating power. However, despite the variation in assignment power between different AFLP primer combinations, the results showed that only one primer combination may be sufficient to reach an assignment success of 95% at ε = 0, and that such a high level of assignment success was maintained at high stringency level (ε = 3) when using the three primer sets. The choice of the stringency level to be used with assignment methods depends mainly on a trade-off between performance of the assignment procedure (in terms of overall proportion of individuals assigned) and confidence level over assignments. In this study, AFLPs have proved to be powerful with a fairly small increase in proportion of failures, even at high stringency level (ε = 3) when using all three primer combinations. This suggests that, with AFLPs, the trade-off between confidence and assignment performance may be small when using a sufficient number of loci. Consequently, a high confidence level can be achieved even when one wants to favour assignment performance.

Effect of number of loci on assignment success

Figure 6 shows two panels of curves representing assignment success as a function of minimal frequency differential. For all primer combinations and stringency levels, the assignment success was maximal when using all informative loci (*FD* > 0; left-hand side of Fig. 6). When the number of loci was decreased gradually to retain loci with higher frequency differentials the assignment success decreased slowly at ε = 0, whereas it decreased quickly at ε = 3, with the exception of GGTG or when the three primer combinations were pooled together. With GGTG and the three primer combinations, respectively, 94.7% and 90% of the maximal assignment success were obtained using only six loci with frequency differentials higher than 0.70 at ε = 0. At ε = 3, 78.6% and 68% of the maximal assignment success was still achieved with these six loci. This suggests that loci with high frequency differentials contribute in large part to assignment power. However, the analysis of the effect of the number of loci on assignment success in the whitefish case study revealed that the addition of numerous loci with lower discriminating power enhanced assignment power further, resulting in a higher assignment success when all sources of information were used (all loci with *FD* > 0).

Loci having a frequency differential below a certain threshold (*FD* < 0.50) are discarded commonly because their power to discriminate populations is considered too

Ecotype	Bwf1	Bwf2	C2-157	C4-157	Cocl22	Cocl23	A_T	H_M
Dwarf								
N	39	40	40	39	39	40		
A	8	6	9	7	6	5	41	
A_C	220	157	147	293	125	266		
F_C	0.61	0.55	0.39	0.68	0.55	0.63		
A_R	212–224	147–159	121–167	285–301	105–127	260–270		
H_O	0.60	0.63	0.76	0.52	0.60	0.56		0.61
H_E	0.62	0.67	0.63	0.30	0.89	0.59		0.62
Normal								
N	40	40	40	35	37	40		
A	6	4	10	5	6	6	38	
A_C	220	157	145	293	123	266		
F_C	0.61	0.91	0.29	0.36	0.46	0.55		
A_R	212–224	147–159	121–167	285–301	105–127	260–270		
H_O	0.57	0.16	0.82	0.76	0.64	0.58		0.59
H_E	0.51	0.15	0.63	0.66	0.76	0.58		0.55

Table 2 Allelic variability at six microsatellite loci from sympatric lake whitefish ecotypes in East lake. These data are reported from Lu & Bernatchez (1999). Number of samples used for genetic analysis (*N*), number of alleles at each locus (*A*), total number of alleles at six loci (A_T), most common alleles (A_C; in base pairs), frequencies of the most common alleles (F_C), range of allele size (A_R), observed heterozygosity (H_O) and gene diversity (H_E) at each locus, and mean within-ecotype heterozygosity at six loci (H_M)

small (e.g. Shriver *et al*. 1997). As pointed out by Smouse & Chevillon (1998), loci with low frequency differentials may be more prone to sampling error. In accordance, the conditions that produced the largest loss of probability of correct assignment due to sampling error in our analytical investigation were small frequency differentials, small sample sizes and *C*-loci ($f_A = 0.25$). However, in all but the worst conditions (sample size $n = 10$), it was possible to reach the targeted 90% probability of correct assignment by increasing the number of loci with low frequency differentials. Furthermore, with AFLP, the likelihood of obtaining loci with low frequency differential under weak population differentiation may be counterbalanced by an increased probability of finding loci with high frequency differentials due to the large number of loci generated. Consequently, it is most likely that, under empirical conditions, the sets of loci will contain few highly informative loci (high *FD*) and many least informative (low *FD*) loci. As can be seen from the case study results the use of such sets, without discarding loci with low frequency differential values, is likely to maximize assignment success. In the advent of a complete absence of highly informative loci it would remain possible, on the basis of our analytical results, to achieve high assignment success by using enough loci with low frequency differentials and adequate sample sizes.

AFLPs vs. microsatellites

Moderate to high levels of genetic diversity were observed at microsatellite loci within each sample, with the number of alleles per locus varying from 5 to 10, and gene diversity estimates varying between 0.16 and 0.82 (Table 2). The average heterozygosity (H_M) across loci in the dwarf and normal ecotype were similar. Homogeneity tests of allele frequency distribution showed that dwarf and normal whitefish ecotypes were differentiated genetically ($P < 0.05$) and the extent of genetic differentiation between both ecotypes was low ($\theta = 0.058$, SD = 0.013–0.110).

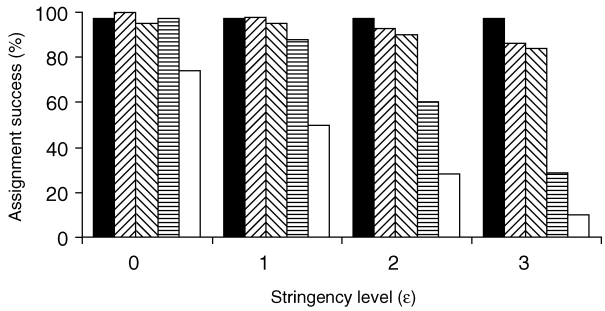

Fig. 7 Comparison of the assignment success between microsatellite and AFLP at four levels of stringency ($\varepsilon = 0$, $\varepsilon = 1$, $\varepsilon = 2$ and $\varepsilon = 3$). From left to right, bars in each histogram refer to the three primer combinations combined, CATA, GGTG, CCTC and microsatellite loci, respectively.

The 172 AFLP loci were more powerful than the six microsatellite loci in discriminating the source of an individual among putative populations. Indeed, AFLP markers resulted consistently in higher assignment success than microsatellite loci at all levels of stringency, regardless of the AFLP primer combination used (Fig. 7). Furthermore, the log-likelihood differences between dwarf and normal populations obtained with AFLP for each individual were generally much larger than those obtained with the six microsatellites (Fig. 8). In only one individual (1.7%) did the log-likelihood difference fall below the $\varepsilon = 3$ level with AFLP in contrast to 52 individuals (90%) with

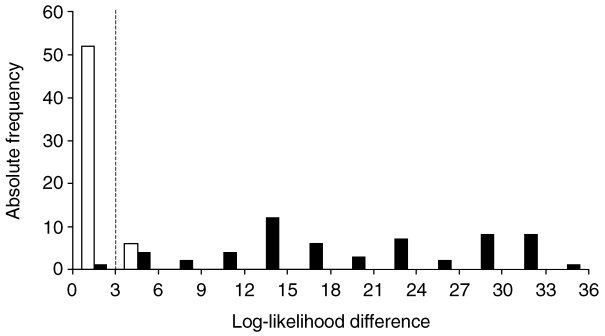

Fig. 8 Histogram of log-likelihood differences (absolute difference between the likelihood of an individual being assigned to the dwarf population and to the normal population) obtained from AFLP (■) and microsatellite (□) markers. The dashed line indicates our highest level of stringency (ε = 3).

microsatellites. Comparable low assignment power with microsatellites was obtained in other studies showing weak population structuring (e.g. Roques *et al*. 1999; Ruzzante *et al*. 2001). Low assignment power with microsatellites in these situations was most probably a direct consequence of using an insufficient number of loci, given the small allele frequency differentials among weakly structured populations.

Given the relative ease with which large number of AFLP markers can be obtained without any genetic information on the species, this study showed that AFLPs may be an excellent alternative to microsatellites in order to enhance resolution in studies of population assignment, especially when population differentiation is weak. On the other hand, AFLPs have some significant drawbacks, the most significant of which is that their dominant nature makes it impossible to evaluate departure from Hardy–Weinberg equilibrium. The fact that band homology is most often assumed instead of being demonstrated from sequence analysis may also hamper their reliability. Furthermore, AFLPs may not be suitable for all population assignment applications. For example, they may not provide enough information to detect immigrant individuals directly, unlike methods that use codominant markers (Rannala & Mountain 1997), although this remains to be assessed rigorously. Consequently, for some research programs involving large-scale description of the population structure of a species, the advantages of AFLPs for population assignment may be offset by some of the long-term benefits of developing microsatellites. Indeed, microsatellites markers may be useful for studies involving multiple species over a long period of time without the possible homology problems. Also, due to their codominance, they provide more information on inbreeding coefficients. In summary, in order to make the appropriate choice of marker, one should compare the benefits and drawbacks of each type of loci relative to the projects that are undertaken.

Acknowledgements

We are grateful to Lucie Papillon and Robert St-Laurent for technical assistance in the laboratory, and to Sean Rogers for his constructive comments on the manuscript. We are also grateful to two anonymous reviewers for their helpful comments on a previous version of the manuscript. This work was supported by the Canadian Research Chair in conservation genetics of aquatic organisms and a research grant from the Natural Science and Engineering Council (NSERC, Canada) to LB. D. Campbell is supported by an NSERC Postgraduate Scholarship.

References

Albertson RC, Markert JA, Danley PD, Kocher TD (1999) Phylogeny of a rapidly evolving clade: the cichlid fishes of lake Malawi, East Africa. *Proceedings of the National Academy of Sciences USA*, **96**, 5107–5110.

Bensch S, Helbig AJ, Salomon M, Seibold I (2002) Amplified fragment length polymorphism analysis identifies hybrids between two subspecies of warblers. *Molecular Ecology*, **11**, 473–481.

Bernatchez L, Duchesne P (2000) Individual-based genotype analysis in studies of parentage and population assignment: how many loci, how many alleles? *Canadian Journal of Fisheries and Aquatic Sciences*, **57**, 1–12.

Congiu L, Dupanloup L, Patarnello T *et al.* (2001) Identification of interspecific hybrids by amplified fragment length polymorphism: the case of sturgeon. *Molecular Ecology*, **10**, 2355–2359.

Cornuet JM, Aulagnier S, Lek S, Franck P, Solignac M (1996) Classifying individuals among infra specific taxa using microsatellite data and neural networks. *Conseil de Recherche Académique Des Sciences de la Vie*, **319**, 1167–1177.

Cornuet JM, 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.

Dallimer M, Blackburn C, Jones PJ, Pemberton JM (2002) Genetic evidence for male biased dispersal in the red-billed quelea *Quelea quelea*. *Molecular Ecology*, **11**, 529–533.

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

Duchesne P, Bernatchez L (2002) AFLPOP: a computer program for simulated and real population allocation based on AFLP data. *Molecular Ecology Notes*, **2**, 380–383.

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.

Ellegren H, Savolainen P, Rosen B (1996) The genetical history of an isolated population of the endangered grey wolf *Canis lupus*: a study of nuclear and mitochondrial polymorphisms. *Philosophical Transactions of the Royal Society of London B*, **351**, 1661–1669.

Estoup A, Anger B (1998) Microsatellites and minisatellites for molecular ecology: theoretical and experimental considerations. In: *Advances in Molecular Ecology* (ed. Carvhalo GR), pp. 55–86. IOS Press, The Netherlands.

Estoup A, Garnery L, Solignac M, Cornuet JM (1995) Microsatellite variation in honey bee (*Apis melifera* L.) populations: hierarchical genetic structure and test of the infinite allele and stepwise mutation models. *Genetics*, **140**, 679–695.

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

Haig SM (1997) Population identification of western hemisphere shorebirds throughout the annual cycle. *Molecular Ecology*, **6**, 412–427.

Hansen MM (2002) Estimating the long-term effects of stocking domesticated trout into wild brown trout (*Salmo trutta*) populations: an approach using microsatellite DNA analysis of historical and contemporary samples. *Molecular Ecology*, **11**, 1003–1016.

Hansen MM, Kenchington E, Nielsen EE (2001) Assigning individual fish to populations using microsatellite DNA markers. *Fish and Fisheries*, **2**, 93–112.

Lu G, Bernatchez L (1999) Correlated trophic specialization and genetic divergence in sympatric lake whitefish ecotypes (*Coregonus clupeaformis*): support for the ecological speciation hypothesis. *Evolution*, **53**, 1491–1505.

Mountain JL, Cavalli-Sforza LL (1997) Multilocus genotypes, a tree of individuals, and human evolutionary history. *American Journal of Human Genetics*, **61**, 705–718.

Mueller UG, Wolfenbarger LL (1999) AFLP genotyping and fingerprinting. *Trends in Ecology and Evolution*, **14**, 389–394.

Neff BD, Gross MR (2001) Microsatellite evolution in vertebrates: inference from AC dinucleotide repeats. *Evolution*, **55**, 1717–1733.

Nielsen EE, Hansen MM, Loeschcke V (1997) Analysis of microsatellites DNA from old scale samples of Atlantic salmon *salmo salar*: a comparison of genetic composition over 60 years. *Molecular Ecology*, **6**, 487–492.

Orgen R, Thorpe RS (2002) The usefulness of amplified fragment length polymorphism markers for taxon discrimination across graduated fine evolutionary levels in Caribbean *Anolis* lizards. *Molecular Ecology*, **11**, 437–445.

Paetkau D, Calvert W, Stirling I, Strobeck C (1995) Microsatellite analysis of population stucture 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.

Palsboll PJ, Allen J, Berube M *et al.* (1997) Genetic tagging of humpback whales. *Nature*, **388**, 767–769.

Potvin C, Bernatchez L (2001) Lacustrine spatial distribution of landlocked Atlantic salmon populations assessed across generations by multilocus individual assignment and mixed-stock analyses. *Molecular Ecology*, **10**, 2375–2388.

Primmer CR, Koskinen MT, Piironen J (2000) The one that did not get away: individual assignment using microsatellite data detects a case of fishing competition fraud. *Proceedings of the Royal Society of Biological Sciences, Series B*, **267**, 1699–1704.

Rannala B, Mountain JL (1997) Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences USA*, **94**, 9197–9221.

Rieseberg LH (1998) Genetic mapping as a tool for studying speciation. In: *Molecular Systematics of Plants* (eds Soltis PS, Doyle JJ), pp. 459–487. Chapman & Hall, New York.

Rogers SM, Campbell D, Baird SJE, Danzmann RG, Bernatchez L (2001) Combining the analyses of introgressive hybridisation and linkage mapping to investigate the genetic architecture of population divergence in the lake whitefish (*Coregonus clupeaformis*, Mitchill). *Genetica*, **111**, 25–41.

Roques S, Duchesne P, Bernatchez L (1999) Potential of microsatellites for individual assignment: the North atlantic redfish (genus *Sebastes*) species complex as a case study. *Molecular Ecology*, **8**, 1703–1717.

Ruzzante DE, Hansen MM, Meldrup D (2001) Distribution of individual inbreeding coefficients, relatedness and influence of stocking on native anadromous brown trout (*Salmo trutta*) population structure. *Molecular Ecology*, **10**, 2107–2128.

Shriver MO, Smith MW, Jin L (1997) Ethnic affiliation estimation by use of population-specific DNA. *American Journal of Human Genetics*, **60**, 957–964.

Smouse PE, Chevillon C (1998) Analytical aspects of population-specific DNA fingerprint for individuals. *Journal of Heredity*, **89**, 143–150.

Smouse PE, Spielman RS, Park MH (1982) Multiple-locus assignment of individuals to groups as a function of the genetic variation within and differences among human populations. *American Naturalist*, **119**, 445–463.

Vos P, Hogers R, Bleeker M *et al.* (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research*, **23**, 4407–4414.

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

Wilding CS, Butlin RK, Grahame J (2001) Differential gene exchange between parapatric morphs of *Littorina saxatilis* detected using AFLP markers. *Journal of Evolutionary Biology*, **14**, 611–619.

Young WP, Ostberg CO, Keim P, Thorgaard GH (2001) Genetic characterization of hybridization and introgression between anadromous rainbow trout (*Oncorhynchus mykiss irideus*) and coastal cutthroat trout (*O-clarki clarki*). *Molecular Ecology*, **10**, 921–930.

This study is part of D. Campbell's MSc thesis on the genetic basis of reproductive isolation in the lake whitefish (*Coregonus clupeaformis*). Pierre Duchesne is a mathematician working in L. Bernatchez's group. His main interest is in the improvement of methods for parentage and population assignments. He is also interested in investigating the effect of supportive breeding programs on the genetic integrity of natural populations. The major interests of L. Bernatchez are in the understanding of the patterns and processes of molecular and organismal evolution, as well as their significance to conservation.

Appendix I

Expectancy of P within the two-populations model

We are given fixed values of l, the number of loci, and (f_A, f_B), the presence frequencies in populations A and B, constant across all loci. Let us assume that G_A is a genotype from A. We define S as the sum of the presences over all loci of G_A. The computation of P, the expected number of correct assignments over all decisions [assignments, both correct and incorrect, and failures (no assignment)] may be divided into four steps:

(i) P*(S): probability of* S *presences within A* At each of its loci, G_A has probability f_A to show a presence (1). Each presence may be viewed as a success in a Bernoulli trial. Therefore, the sum of presences S across the l loci of G_A has a binomial distribution. Accordingly, the probability that G_A contains $S = 0, 1, 2, 3, \ldots l$ presences is precisely:

$$p(S) = C(l,S)\, f_A^S\, (1 - f_A)^{l-S} \quad S = 0, 1, 2, 3, \ldots l$$

where $C(l, S)$ stands for the number of possible choices of S items (loci) among l.

(ii) Log-likelihoods of S *presences* For each value of S, the two log-likelihoods (one in each population) are:

$$\log L_A(S) = S \log(\phi_A) + (l - S) \log(1 - \varphi_A)$$

and

$$\log L_B(S) = S \log(\varphi_B) + (l - S) \log(1 - \varphi_B)$$

where ϕ_A, ϕ_B are the estimates of f_A, f_B.

Because log(0) is not defined, it is necessary to replace frequencies of zero ($\phi = 0$) by some appropriate value. We chose $\phi = 1/(n + 2)$ after having run simulations to compare this with smaller replacement values such as 0.001 and also to $\phi = 1/(n + 1)$. We found that $\phi = 1/(n + 2)$ produced consistently better results in terms of assignment efficiency (data not shown).

(iii) Assignment decision The assignment decision is based on the difference in log-likelihoods:

$$DL(S) = \log L_A(S) - \log L_B(S)$$

At this point we have pairs ($DL(S)$, $p(S)$) for each $S = 0, 1, 2, 3, \ldots l$

Given assignment criterion ε, the three possible decisions for G_A with S presences are:

assignment to A if $DL(S) > \varepsilon$
assignment to B if $DL(S) < -\varepsilon$
no assignment (failure) if $-\varepsilon \le DL(S) \le \varepsilon$

(iv) Probability of correct assignment of G_A *and overall* P

The probability P_A of correct assignment of genotypes from A is the sum of all $P(S)$ such that $DL(S) > \varepsilon$. Alternatively, the probability of misassignment of genotypes from A would be the sum of all $P(S)$ such that $DL(S) < -\varepsilon$. The probability P_B of correct assignment of genotypes belonging to population B can be computed in a similar fashion.

Finally, the probability of correct assignment of genotypes from A or B is:

$P = pr_A P_A + pr_B P_B$ where
pr_A, pr_B = *proportion of A, B specimens*

Because we assumed no prior knowledge on the relative sizes of the populations we used $pr_A = pr_B = 0.5$ in all calculations. The criteria parameter ε was set to 0. We assumed that ϕ_A, $\phi_B = f_A$, f_B except in the Effects of sampling errors on the probability of correct assignment section to simulate errors on presence frequency estimates.

Appendix II

Expectancy of P *under sampling error model*

(i) Sampling procedure of population presence frequencies (f_s*)* Here, we allow the possibility for sampling error so that f_i, the population frequency presence on locus i, may be viewed as a random variable. We denote NP_i the number of presences on locus i computed over all sample specimens. The size of the population is assumed to be very large compared to the sample size n_i so that NP_i is considered to have binomial distribution. Keeping in mind that NP_i and f_i relate to a specific locus, we now drop the locus index so as to simplify all subsequent expressions. The probability (likelihood) that NP presences among n specimens are observed on a locus with population frequency f:

$$L(NP \mid n,f) = C(n,NP)\, f^{NP}\, (1 - f)^{n-NP}$$

Because the sample size n and the number of presences NP are known quantities, L($NP \mid n$, f) is a polynomial function in f, for intance P(f). In order to obtain the probability density of f, for instance D(f), we re-scale P(f) so that $\int D(f) = 1$. Hence:

$$D(f) = \frac{P(f)}{\int_0^1 P(f)}$$

To sample f from D(f) we use the 'inverse' method. This method requires to compute the cumulative density function of f:

$$C(f) = \int_0^f D(f)df - \int D(f)df \Big|_0$$

Sampling from D(f) is performed in two steps. First a number, for instance X, is generated from a uniform random generator U(0,1). Then X is transformed into:

$$f = C^{-1}(X)$$

where C^{-1} is the inverse cumulative density of f. The standard Maple numerical solver was used to solve for f (from X).

(ii) Sampling procedure from P In the two-populations (analytical) model, the true frequencies (f_A, f_B) were held constant across all loci and no sampling error occurred. To assess the loss of allocation power due to sampling error on frequencies, we now hold the frequency estimates (ϕ_A, ϕ_B), i.e. NP_A, NP_B constant across loci. However, f_A, f_B have become random variables with probability densities D(f_A), D(f_B) computed as above. Clearly, the proportion of correct allocations *P* has also become a random variable. To obtain estimates of the expectancy of *P* and so be able to assess the loss of power resulting from sampling error, we have to sample from *P*. We describe below the procedure to sample a single value from *P*. The parameters of the procedure are the sample size *n*, the number of presences NP_A, NP_B, the number of loci *l* and the number of random (artificial) genotypes NG. Based on *n* and NP_A, generate an array $\mathbf{f_A}$ of *l* random frequencies using the above sampling procedure for f. Based on *n* and NP_B, generate an array $\mathbf{f_B}$ of *l* random frequencies. Populations A and B are now, respectively, described by arrays $\mathbf{f_A}$, $\mathbf{f_B}$.

1. Following the distribution of presences at each locus, generate NG random A-genotypes from $\mathbf{f_A}$. Similarly, generate NG B-genotypes from $\mathbf{f_B}$.
2. For each A-genotype, compute log-likelihoods within A and B. Assign to population of largest log-likelihood following the same procedure as in the two-populations (analytical) model. Similarly, compute log-likelihoods and assign B-genotypes.
3. Compute the proportion of correctly assigned A- and B-genotypes. This proportion is one random value sampled from *P*.

*(iii) Computation of E(*P*)* To estimate E(*P*) from frequency estimates ϕ_A, ϕ_B, we generated 100 frequency array samples ($\mathbf{f_A}$, $\mathbf{f_B}$). Each array sample represented one possible pair of populations (A, B). With each pair of populations, 1000 A-genotypes and 1000 B-genotypes were generated and assigned. Hence, a total of 200 000 genotypes served as basis to estimate E(*P*) as a function of sample size *n* and presence frequency estimates ϕ_A, ϕ_B. E(*P*) is the average of 100 random values sampled from *P*.