Genetica (2005) 00:1–12
DOI 10.1007/s10709-005-5365-9

Effects of coancestry on accuracy of individual assignments to population of origin: examples using Great Lakes lake trout (*Salvelinus namaycush*)

Bruno Guinand[1,2], Kim T. Scribner[1], Kevin S. Page[1,4], Kristi Filcek[1], Laura Main[1] & Mary K. Burnham-Curtis[3]

[1]*Department of Fisheries and Wildlife, Michigan State University, East Lansing, MI 48824-1222, USA;* [2]*UMR CNRS 5171 Génome, Populations, Interactions, Adaptation, Station Méditerranéenne de l'Environnement Littoral – Quai de la Daurade, 34200, Sète, France; (Phone: +33-4-67-46-33-82, Fax: +33-4-67-46-33-99, E-mail: guinand@univ-montp2.fr);* [3]*National Forensics Laboratory, U.S. Fish and Wildlife Service, 1450 Main St., Ashland, OR USA;* [4]*Minnesota Department of Natural Resources, 1601 Minnesota Dr., Brainerd, MN 56401-3971, USA*

Received 6 June 2005 Accepted 22 November 2005

Key words: assignment test, coancestry, gene correlations, inbreeding, lake trout

Abstract

Methods for assigning individuals to population of origin are widely used in ecological genetics, resources management, and forensics. Characteristics of genetic data obtained from putative source populations that enhance accuracy of assignment are well established. How non-independence within and among unknown individuals to be classified [i.e., gene correlations within individuals (inbreeding) and gene correlations among individuals within groups (coancestry)] affect assignment accuracy is poorly understood. We used empirical data for six microsatellite loci and offspring from full-sib crosses of hatchery strains of lake trout (*Salvelinus namaycush*; Salmonidae) representing known levels of coancestry (mean $\theta = 0.006$ and 0.06) within families to investigate how gene correlations can affect assignment. Additional simulations were conducted to further investigating the influence of allelic diversity (2, 6 or 10 alleles per locus) and inbreeding ($F = 0.00$, 0.05, and 0.15) on assignment accuracy for cases of low and high inter-population variance in allele frequency (mean $F_{st} = 0.01$ and 0.1, respectively). Inbreeding had no effect on accuracy of assignments. In contrast, variance in assignment accuracy across replicated simulations, and for each empirical case study increased with increasing coancestry, reflecting non-independence of probabilities of correct assignment among members of kin groups. Empirical estimates of assignment error rates should be interpreted with caution if appreciable levels of coancestry are suspected. Additional emphasis should be placed on sampling designs (spatially and temporally) that define or minimize the potential for sampling related individuals.

Introduction

Fundamental understanding of ecological and genealogical relationships between populations is a prerequisite for effective population management and conservation. Allele frequencies estimated using highly polymorphic DNA markers such as microsatellite loci are widely recognized as a viable means to define population boundaries, and to estimate rates of gene flow among populations (Waser & Strobeck, 1998; Luikart & England, 1999; Manel, Gaggiotti & Waples, 2005). Statistical tools such as assignment tests have been widely used to place individuals to putative populations of origin when spatial genetic variation among populations exists. Assignment tests have been used widely in

Journal : GENE | Dispatch : 8-12-2005 | Pages : 12
CMS No. : DO00025365 | ☐ LE | ☐ TYPESET
MS Code : GENE150R2 | ☑ CP | ☑ DISK
PDF-OUTPUT

many ecological and evolutionary contexts, including applied fisheries science (Hansen, Kenchington & Nielsen, 2001a), to establish relationships among individuals within and among groups (e.g., Roques, Duchesne & Bernatchez, 1999; Koskinen, Piironen & Primmer, 2001; Nielsen et al., 2001), identification of introgression and hybrid individuals (Martinez et al., 2001), and ecotypes (Taylor et al., 2000), forensics (Primmer, Koskinen & Piironen, 2000), phylogeographical analyses (King et al., 2001), documenting contributions of stocked individuals to natural populations or evaluation of supportive breeding programs (Hansen et al., 2000, 2001b), and to infer rates of dispersal (Berry, Tocher & Sarre, 2004; Castric & Bernatchez, 2004). Empirical and theoretical studies have shown assignment tests to be useful for purposes of classification when moderate numbers of loci are employed, characterized by moderate to high numbers of alleles per locus, and when populations are moderately to highly differentiated ($F_{st} \approx 0.05$ or higher; Cornuet et al., 1999; Bernatchez & Duchesne, 2000; Paetkau et al., 2004). Studies have focused on properties of the putative source populations rather than on the unknown individuals to be classified (Manel et al., 2005).

Despite the wide use of assignment tests, several constraints may limit their use. Tests generally assume that source populations are in Hardy–Weinberg equilibrium and loci are independent (i.e., no linkage disequilibrium). Tests assume adequate samples sizes of source populations, and random and equitable sampling across genealogical groups within sources populations to ensure accurate estimation of allele (and expected genotype) frequencies (Guinand et al., 2004). Regardless of the statistical basis of different assignment methods, all assume that the genetic markers employed provide independent information on an individual's ancestry. The importance of violations of assumptions to assignment accuracy have been investigated for source samples (e.g., linkage in admixed populations; Falush, Stephens & Pritchard, 2003), but are likely to be similarly important for samples of unknowns to be classified. For example, if unknown individuals collected from the same locale were related (i.e., gene correlations or non-zero coancestry *between* individuals within the same breeding group; Wright, 1969; Chesser, 1991a, b), non-independence due to shared pedigree could influence population assignment. Individuals characterized by non-zero levels of coancestry could be sampled from populations of low effective size, if the variance in adult reproductive success is high, or due to other life-history and behavioral factors leading to kin-structured populations (Sugg et al., 1996).

Levels of inter-population variance in allele frequency (F_{st}; Wright, 1965) have been widely used to predict the accuracy with which individuals can be assigned to population of origin. Fixation indices (*F*-statistics; Wright, 1965) may also be interpreted in light of non-independence among individuals, based on measures of inbreeding (F), coancestry (θ) and inter-group correlations (α) (Wright, 1969; Cockerham, 1973; Chesser, 1991a, b). As correlations increase, less genetic variation is apportioned at other levels of population structure (Chesser, 1991a, b). Hence, increasing coancestry leads to lower variance within populations, and concomitantly to reapportionment of greater portions of overall variance among families within populations and among populations. Understanding how the relative apportionment of genetic variance influences assignment accuracy has not been rigorously examined. For example, Cornuet et al. (1999) performed extensive simulations focusing on varying levels of population differentiation (F_{st}), but not on other components of genetic variance such as inbreeding (F) and coancestry (θ).

Using empirical data sets, we investigate how inbreeding and coancestry of unknown individuals to be classified influences assignment accuracy. Specifically, we created mixtures composed of individuals from different strains of lake trout (*Salvelinus namaycush*; Salmonidae), where varying proportions of individuals from each strain were related at the level of full-sibs. Simulations were also conducted on a range of values for parameters including the number of loci, number of alleles per locus, and level of population differentiation, that reflect characteristics observed for empirical data sets typically studied in nature. We demonstrate that coancestry, but not inbreeding levels of unknown individuals to be classified influenced accuracy of assignment decisions.

Materials and methods

Three levels of mean coancestry were considered: no coancestry ($\theta_j = 0.00$), low coancestry ($\theta_j = 0.006$), and high coancestry ($\theta_j = 0.06$). All

Journal : GENE | Dispatch : 8-12-2005 | Pages : 12
CMS No. : DO00025365 | ☐ LE | ☐ TYPESET
MS Code : GENE150R2 | ☑ CP | ☑ DISK

analyses were conducted using source populations characterized by empirically estimated variance in allele frequency ($F_{st} \approx 0.01$ and $F_{st} \approx 0.10$). In all cases, accuracy of assignment was computed by assigning individuals of a given coancestry level to source populations. Genotype frequencies in all simulated and empirical source populations conformed to Hardy–Weinberg expectations (i.e., no significant coancestry).

Empirical samples

Three lake trout hatchery strains were selected for empirical evaluations. The Marquette (SMD) and Seneca (SLW) strains represented a pair-wise comparison where allele frequencies were highly divergent, and represented a high F_{st} case ($F_{st} \approx 0.10$). The SMD and Isle Royale (SIW) strains were selected as the low F_{st} case ($F_{st} \approx 0.01$). Details on hatchery strains may be found in Page et al. (2003). Spawning of adults were conducted by U.S. Fish and Wildlife Service personnel during annual lake trout egg takes for each domestic broodstock during the fall of 1998. All crosses for each family involved one male and one female. Fertilized eggs from each full-sib family were incubated individually and upon emergence were placed together in 95% non-denatured ethanol for analysis.

Reproductive data including the total number of lake trout pairs spawned and total numbers of juveniles produced from each mating were used to calculate average levels of coancestry (mean θ; Chesser, 1991a). We sampled 15 offspring from each of four full-sib families and 6 individuals from each of 10 full-sib families to generate empirical sample sets with high and low levels of coancestry, respectively for each strain comparison in our analyses. Coancestry represents correlations of genes between individuals within the same family group, describing the probability of identity by descent for two alleles drawn randomly from each of two individuals. Coancestry values between full-sibs were 0.25. Gene correlation matrices were derived for individual families as described by Chesser (1991a). The number of juveniles sampled per family ($b = 6$ or 15), number of families ($n = 10$ or 4), and total numbers of individuals within a strain used in the experiment (N) were used to estimate the average coancestry of the j^{th} strain as per Chesser (1991a):

$$\theta_j = \frac{\sum_{i=1}^{n} b_i^2 - b_i}{4(N^2 - N)} \quad (1)$$

Using Equation (1) we selected full-siblings from families to empirically establish groups with known high ($\theta_j = 0.06$) and low ($\theta_j = 0.006$) empirical levels of known coancestry using offspring from four and 10 family crosses, respectively. Baseline SIW, SLW and SMD broodstocks were used as source populations in assignments where θ for each source was assumed to be zero.

Laboratory analyses of empirical samples

DNA extraction of tissues from emergent juveniles ($N = 382$) was performed using a proteinase K digestion and a modified Puregene extraction protocol (Gentra, Inc., Minneapolis, MN). DNA was resuspended in 50 μl of TE buffer (10 mM Tris–HCL, pH 8.0, 1 mM EDTA). Fluorometry was used to determine DNA concentrations. One hundred nanograms of DNA was used for each PCR. We used six microsatellite loci that were originally developed for other salmonid species including brook trout (*Salvelinus fontinalis*) (*Sfo1*, *Sfo12* and *Sfo18*; Angers, Bernatchez, 1996), pink salmon (*Onchorynchus gorbuscha*) (*Ogo1a*; Olsen, Bentzen & Seeb, 1998), bull trout (*Salvelinus confluentus*) (*Scoμ19*; Taylor et al., 2001), and Atlantic salmon (*Salmo salar*) (*Ssa85*; O'Reilly et al., 1996). PCRs were performed in 25 μl volumes using conditions provided by the respective authors. PCR products were screened using 6% polyacrylamide vertical gels. Products were visualized by a Hitachi FMBIO II Multi-View scanner and associated software. Microsatellite fragments were sized manually using a 20 bp internal lane standard. Several individuals of known genotype were used as additional allele size standards on each gel.

Summary statistics for empirical data

Estimates of allele frequency for the baseline SMD, SIW, and SLW broodstocks were obtained independently (Page, 2001) for each of 6 microsatellite loci based on >60 adults per broodstock.

Journal : GENE | Dispatch : 8-12-2005 | Pages : 12
CMS No. : DO00025365 | ☐ LE | ☐ TYPESET
MS Code : GENE150R2 | ☑ CP | ☑ DISK

Previous simulation studies (Guinand et al., 2004) have shown that sample sizes at or exceeding this level are sufficient to accurately estimate allele frequencies for source populations under all conditions evaluated in this study. Estimates of deviations of observed genotypic proportions from Hardy–Weinberg expectations were made using the Markov Chain method of Guo and Thompson (1992) implemented in the program Genepop (Raymond & Rousset, 1995). *F*-statistics (Weir & Cockerham, 1984) describing levels of allelic variance among individuals within and among strains were estimated using Genetix 4.02 (Belkhir et al., 2001). Statistical significance for all Hardy–Weinberg tests and tests of inter-strain differences in allele frequency were based on probability levels adjusted for multiple comparison tests using sequential Bonferroni tests (Rice, 1989). Evidence for linkage disequilibrium was tested using Genetix 4.02.

Assignment tests

We first employed the likelihood-based assignment test of Paetkau et al. (1995) to assign individuals to strain of origin. This approach was used primarily due to the wide use of likelihood-based methods in the empirical literature. Our working hypothesis was that genetic correlations among individuals would lead to greater probabilities of concurrent correct or misclassification of members of the same family group. We then summarized results characterizing the importance of gene correlations and apportionment of genetic variation on assignment error rate. We estimated accuracy of individual strain classification using multilocus genotype frequencies of each lake trout broodstock as described previously (Page, 2001; Page et al., 2003) using the leave-one-out procedure (Efron, 1983). We assigned progeny from low coancestry and high coancestry groups of unknowns for each inter-strain comparison to determine whether accuracy of individual assignments was random across samples. To investigate whether assignment accuracy covaried non-independently among individuals as a function of familial relationship (i.e., a 'family effect' characterized by elevated inter-individual gene correlations), we conducted chi-square tests to examine whether progeny from the same parental cross were more likely to be assigned correctly to source populations relative to random members of the entire sample.

We also used the Bayesian method of Pritchard et al. (2000), implemented in the program Structure (v. 1.0). The program uses multilocus genotypes to infer population structure and to assign individuals based on posterior probabilities to populations (or strains). Results were based on 100,000 or 500,000 Markov Chain Monte Carlo iterations following a burn-in period of 20,000 or 100,000 iterations for the high- and low F_{st} cases, respectively. For each F_{st} level and each coancestry case, we estimated individual admixture proportions ($\hat{q}$, posterior probability of assignment), representing the estimated proportion of an individual's genotype originating from either parental population. We ranked estimated $\hat{q}$ values from the smallest to the highest value to obtain a distribution of individual admixture proportions for each F_{st} level and each coancestry case as described by Nielsen et al. (2003). We then used a Kolmogorov–Smirnov two-sample test to evaluate pair-wise differences between observed distributions of individual admixture proportions under two coancestry levels. This test compares distributions of $\hat{q}$ values between different groups characterized by different mean coancestry levels. Specifically, pair-wise comparisons were conducted between distributions of individual admixture proportions with no coancestry [mean $\theta_j = 0.00$], low coancestry [mean $\theta_j = 0.006$], and high coancestry [mean $\theta_j = 0.06$]. Test were conducted were considered for each level of empirically estimated variance in allele frequency [$F_{st} \approx 0.01$ and $F_{st} \approx 0.1$]).

Computer simulations

We used computer simulations for the range of parameters represented in the empirical lake trout examples. We simulated population data for 6 loci using three different levels of allelic diversity (2, 6 or 10 alleles per locus). We focused on loci with this range of allelic diversity because individual characterizations will result in higher probabilities of individuals sharing alleles that are identical in state but not identical by descent relative to simulations employing loci with higher allelic diversity. Accordingly, the null hypothesis (no effect of gene correlations on assignment accuracy) would be more difficult to reject. Further, Bernatchez and

Journal : GENE | Dispatch : 8-12-2005 | Pages : 12
CMS No. : DO00025365 | ☐ LE | ☐ TYPESET
MS Code : GENE150R2 | ☑ CP | ☑ DISK

Duchesne (2000) showed that increasing the number of alleles did not significantly improve assignment accuracy.

For simulated data sets, two levels of population differentiation estimated as per Wright (1965) were examined, including a low and high F_{st} case ($F_{st} = 0.01$ and 0.10, respectively). Levels of inter-population variance in allele frequency were consistent with levels of inter-strain differentiation between the low F_{st} and high F_{st} from the empirical lake trout examples, respectively, and reflected ranges of values likely encountered in natural settings. Individuals were simulated by randomly drawing alleles with replacement each allele for each locus, with probabilities associated with multinomial distributions established for each population. Multinomial distributions in each population were characterized by the same number of allelic states. Under conditions of high F_{st} (0.10) and low F_{st} (0.01), allele frequencies for a second population were established as per Wright (1965) with levels of inbreeding or gene correlations within individuals (F) equal to 0.00, 0.05 and 0.15. We assumed no gametic disequilibrium ($\Delta_{AB} = 0$; Weir, 1979). To simulate progeny for use in assignment tests, we randomly selected two individuals as parents and randomly selected one allele per parent per locus to produce progeny genotypes. Progeny arrays were constructed with 60 individuals with contributions from each parental pair proportional to the empirical low and high coancestry levels described above for the empirical lake trout data (i.e., 10 families of 6 individuals in the low coancestry case; 4 families of 15 individuals in the high coancestry case). We conducted 25 replicate simulations for each combination of allelic diversity (2, 6 or 10), coancestry (0.00, 0.006, and 0.06), and inbreeding (0.00, 0.05, 0.15). Inter-population variance in allele frequency was set at $F_{st} = 0.01$ and 0.10. Likelihood-based tests of assignment accuracy were summarized for each combination as the mean and standard error in assignment accuracy over the 25 replicates.

We also computed *mean* distributions of individual admixture proportions over the simulated replicates using the Structure (v. 1.0) program. Characteristics of Markov Chain Monte Carlo and burn in period were identical to the ones used for empirical data sets. We then performed Kolmogorov–Smirnov two-sample tests as described above. For each level of allelic diversity, we made pair-wise comparisons both between different coancestry level for a given inbreeding and F_{st} level, but also between different inbreeding levels for each coancestry and F_{st} level.

Results

Empirical data

No significant linkage disequilibrium was detected in each baseline (broodstock) sample, and genotype frequencies of each broodstock conformed to Hardy–Weinberg expectations. However, increasing coancestry of individuals in groups of unknowns led to significant and increasing departures from HWE expectations for unknown samples (Table 1).

Accuracy of assignment increased with increasing coancestry in both the low and high F_{st} cases (Table 1). Increases in levels of coancestry and lower inter-individual variance within families led to concomitant changes apportionment of genetic variance (greater variance among families and broodstocks), especially when F_{st} was high (Table 1). Higher inter-population variance in allele frequency (F_{st}) led to greater assignment accuracy in all empirical lake trout cases. We found significant differences (Kolmogorov–Smirnov test; $p \leq 0.05$) between distributions of individual admixture proportions for samples from each pair-wise coancestry level comparison in the low F_{st} case (SMD–SIW). No significant difference was detected in the high F_{st} case (SMD–SLW). Results were not significant ($0.05 < p < 0.10$) when comparing $\hat{q}$ distributions between the high coancestry and low coancestry data sets.

To investigate the effects of familial relationship on assignment accuracy, we tabulated the number of offspring from crosses within each strain and full-sib family that were correctly classified to strain of origin. The null hypothesis that assignment accuracy across all individuals was independent of familial origin was rejected. When classifying unknown individuals originating from broodstocks characterized by low inter-population variance in allele frequency (e.g., the low F_{st} case; SIW and SMD lake trout broodstocks; $F_{st} \sim 0.01$) and with both low and high coancestry levels, we found that the majority of individuals from the same families were either correctly or incorrectly

Table 1. Likelihood-based assignment error rates for *Salvelinus namaycush* empirical data sets representing low and high F_{st} cases, and various levels of coancestry

Low F_{st} case	SMD–SIW	Source ($\theta = 0.000$)				Low coancestry (mean $\theta = 0.006$)				High coancestry (mean $\theta = 0.060$)			
		HW	Error rate 35.8%			HW	Error rate 28.3%			HW	Error rate 20.0%		
			F_{it}	F_{is}	F_{st}		F_{it}	F_{is}	F_{st}		F_{it}	F_{is}	F_{st}
Mean number of alleles	5.67	NS	0.002	−0.019	0.013	***	0.147	0.110	0.041	***	0.153	0.108	0.051
SE			0.035	0.031	0.007		0.059	0.054	0.021		0.108	0.109	0.030
High F_{st} case	SMD–SLW	HW	Error rate 9.2%			HW	Error rate 10.6%			HW	Error rate 0.8%		
			F_{it}	F_{is}	F_{st}		F_{it}	F_{is}	F_{st}		F_{it}	F_{is}	F_{st}
Mean number of alleles	5.33	NS	0.086	−0.050	0.129	***	0.183	0.076	0.115	***	0.226	0.044	0.190
SE			0.032	0.028	0.030		0.057	0.065	0.037		0.057	0.086	0.057

NS, not significant; ***$p < 0.001$.
Multilocus results of tests of Hardy–Weinberg (HW) expectations and estimates of fixation indices (and associated standard errors; SE) for low- and high-coancestry levels of unknown individuals are reported. Assignment error rates for source (broodstock) individuals were estimated according to the leave-one-out procedure of Efron (1983).

classified. All four χ^2 tests (2 strains for each of 2 coancestry levels) were significant ($p < 0.05$). For the high F_{st} case (SLW and SMD strains), we found little evidence for lack of independence in assignment as a function of familial membership. Only one of four χ^2 tests were significant for the SLW strain in the low coancestry case ($p = 0.011$). No significant results were detected for the SMD strain for the low ($p = 0.40$) or high ($p = 0.25$) coancestry cases. Since a high proportion of individuals were correctly classified even in the absence of coancestry (Table 1), there was no statistically significant family effect in the high F_{st} case. In contrast, when variance in allele frequency between strains was low (the low F_{st} case), assignment accuracy was greatly affected by coancestry levels of unknown individuals.

Simulated data

Using simulated data we were able to examine effects of coancestry as well as inbreeding on assignment accuracy. Increasing levels of inbreeding and coancestry reflect decreasing variance within families, and concomitantly, a propensity for greater variance *among* families. With increasing inter-population variance in allele frequency (low versus high F_{st}), and when loci with different allelic diversities are used (2, 6 or 10 alleles per locus), there were no effects of inbreeding across the cases considered (Figure 1). Accuracy of assignment was approximately equal and the magnitude of variance across replicates were similar. Assignment error rates and associated standard errors for the low coancestry at $F = 0.01$ and 0.10 were nearly identical across simulations of different levels of allelic diversity (Figure 1).

Contrary to results for inbreeding, the relationship between accuracy of likelihood-based assignment and coancestry was complex using simulated data. Average assignment error rates were generally lower for low coancestry cases compared to the no coancestry cases (except when F_{st} was high for the 6 and 10 allele cases), because assignment accuracy was already high in cases of no coancestry cases (Figure 1(d–f)). Assignment error rates increased for high coancestry cases. Increases in average assignment error rate for higher levels of coancestry was evident for the 6 and 10 allele cases only, especially when variance in allele frequency among source populations (F_{st})

Journal : GENE | Dispatch : 8-12-2005 | Pages : 12
CMS No. : DO00025365 | ☐ LE | ☐ TYPESET
MS Code : GENE150R2 | ☑ CP | ☑ DISK

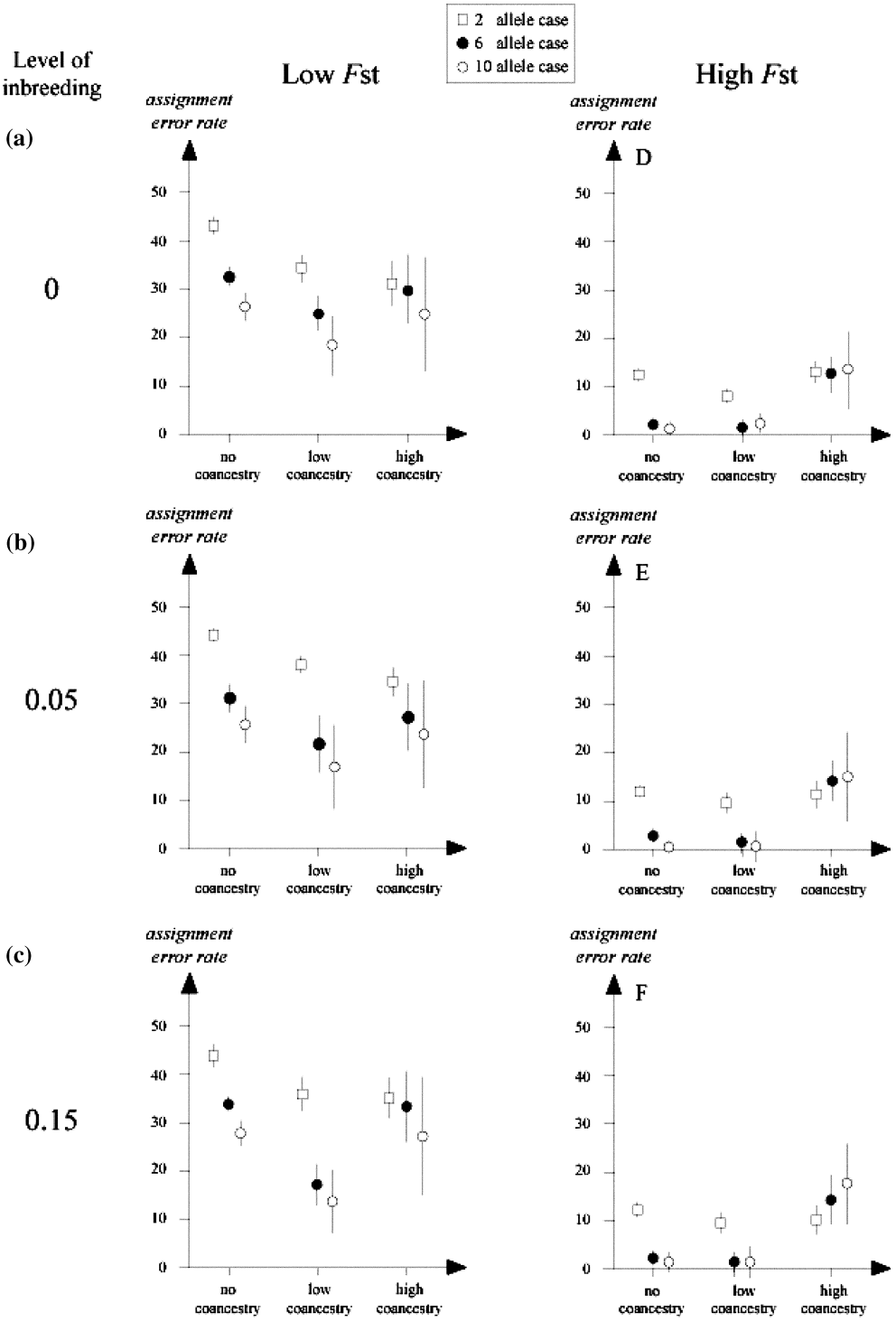

Figure 1. Mean assignment error rates for low- and high-coancestry individuals in *Salvelinus namaycush* empirical data sets representing low and high F_{st} cases. Bars indicate standard errors over replicates.

Journal : **GENE** Dispatch : **8-12-2005** Pages : **12**
CMS No. : **DO00025365** ☐ LE ☐ TYPESET
MS Code : **GENE150R2** ☑ CP ☑ DISK

Table 2. Summary of significant pair-wise changes in mean distributions of individual admixture proportions ($\hat{q}$) over replicated simulated data sets when coancestry (θ) varied

Inbreeding level	Coancestry level	Low F_{st} (0.01)						High F_{st} (0.01)						
		2 alleles		6 alleles		10 alleles			2 alleles		6 alleles		10 alleles	
		Low θ	High θ	Low θ	High θ	Low θ	High θ		Low θ	High θ	Low θ	High θ	Low θ	High θ
0	No θ	*	*	NS	NS	(*)	NS	No θ	NS	NS	NS	(*)	NS	(*)
	Low θ		NS		NS		NS	Low θ		NS		(*)		*
0.05	No θ	*	*	NS	NS	NS	NS	No θ	NS	NS	NS	(*)	NS	(*)
	Low θ		NS		NS		NS	Low θ		NS		*		*
0.15	No θ	*	(*)	*	NS	(*)	NS	No θ	NS	NS	NS	(*)	NS	(*)
	Low θ		NS		(*)		NS	Low θ		NS		(*)		*

NS, not significant.
*$p < 0.05$ when corrected for multiple tests.
(*), $p < 0.05$, not corrected for multiple tests.
Results were provided by Kolmogorov–Smirnov two-sample test.

was high (Figure 1(d–f)). At the high coancestry level, accuracies of assignment in the 2, 6, and 10 allele cases were similar. Generally, no improvement in likelihood-based assignment accuracy was observed when loci characterized by high allelic diversity were used. In simulations of non-zero coancestry, we observed greater variance in assignment accuracy among families and greater inter-replicate variance in assignment accuracy. Ranges of assignment error rates, as assessed by standard errors associated with mean assignment accuracy, increased with level of coancestry for all levels of allelic diversities considered, and for low and high F_{st} cases (Figure 1). This trend was stronger for the 10 allele case (Figure 1).

We used Kolmogorov–Smirnov tests to conduct pair-wise comparisons of individual admixture proportions ($\hat{q}$) for different levels of inbreeding, for levels of allelic diversity and coancestry level. No significant changes associated with level of inbreeding level were observed (results not shown). Conversely, significant changes in distributions of $\hat{q}$ between simulated groups characterized by different coancestry level were observed in both the low and high F_{st} cases for each level of allelic diversity (Table 2).

Discussion

Given the increasing number of molecular genetic studies of kin association and microgeographic genetic structuring (Storz, 1999; Ross, 2001), studies of the effects of non-independence (coancestry) of individuals to be assigned to population of origin on assignment error rates are warranted. The presence of non-zero gene correlations within or among samples of unknown individuals (inbreeding and coancestry, respectively) will reduce variance components within each contributing population as a proportion of the total variance. For fishes, samples characterized by non-zero levels of coancestry and/or inbreeding could be collected if samples are obtained from populations of low effective size, if the variance in female reproductive success is high, if individuals are philopatric and remain in vicinity of hatching location until maturity, or due to group behaviors such as avoidance of predators (Ryman, Allendorf & Ståhl, 1979; Hedgecock, 1994; Hansen, Nielsen & Mensberg, 1997; Gerlach et al., 2001; Bekkevold, Hansen & Loeschke, 2002; Castric et al., 2002). Although demonstration of kin association in fish is debated (Arnold, 2000; Krause et al., 2001; Russell et al., 2004), evidence is accumulating for the importance of non-random genetic structuring and of the occurrence of inbreeding in natural populations (Fontaine & Dodson, 1999; Pouyaud et al., 1999; Gerlach et al., 2001; Castric et al., 2002; Planes & Lenfant, 2002; Planes et al., 2002; Hansen & Jensen, 2005; Kolm et al., 2005; Fraser, Duchesne & Bernatchez, 2005). Ruzzante, Hansen and Meldrup (2001) found that Danish populations of *Salmo trutta* were likely composed of inbred individuals or by mixtures of individuals of

Journal : GENE | Dispatch : 8-12-2005 | Pages : 12
CMS No. : DO00025365 | ☐ LE | ☐ TYPESET
MS Code : GENE150R2 | ☑ CP | ☑ DISK

different demes at scales smaller than sampling locations. Non-independence among individuals to be classified has been cited as potentially influencing individual classification decisions (Hansen et al., 1997; Ruzzante et al., 2001). Using simulations and mixtures of known ancestry, we explicitly address unstudied aspects of how attributes of unknown individuals affect assignment accuracy.

Our empirical and simulation results extend previous empirical observations by examining a range of conditions under which inbreeding and coancestry may be expected to exist in natural populations. In the empirical portion of this study we described results of assignment accuracy based on collections of known full-sibs. Accuracy of assignment was higher when genetic correlations were greater than zero (i.e., with increasing levels of coancestry; Table 1). Using Kilmogorov–Smirnov test, we also demonstrated that changes occurred among distributions of individual admixture proportion (posterior probability, $\hat{q}$) when coancestry levels varied. Based on results of χ^2 tests, we showed that accuracy of assignment was dependent on familial origin, especially in the low F_{st} case (comparison of the SIW and SMD strains). Results for the high F_{st} case (comparison of the SLW and SMD strains) were inconclusive, likely due to low power, because so few individuals were incorrectly classified even in cases of zero coancestry.

The variance among groups (families or populations) is proportional to the coancestry of individuals within groups (i.e., gene correlation between individuals of the same breeding group; Cockerham, 1969, 1973) only in the absence of inbreeding (i.e., gene correlation within individuals) within groups. Hence, because we could not breed related individuals, empirical lake trout data based on full-sib crosses could not simultaneously examine the effects of inbreeding and coancestry on assignment accuracy.

Results from simulations permitted results from empirical cases to be generalized across ranges of parameter values likely to be encountered in natural populations. We found that inbreeding, over the ranges simulated ($F = 0.00$, 0.05, 0.15), did not affect likelihood-based assignment accuracy (Figure 1). No significant changes were observed across distributions of $\hat{q}$ when inbreeding varied, regardless of the level of allelic diversity and coancestry level. Results from simulations also indicated that mean likelihood-based assignment error rates will be generally lower with increasing levels of coancestry among unknown individuals, particularly when coancestry is low. On average, likelihood-based assignment accuracy was minimally effected when coancestry was high (Figure 1). We also documented greater variance in likelihood-based assignment error rates regardless of the level of F_{st} among source populations (Figure 1). Results of Kolmogorov–Smirnov two sample tests further demonstrated that distributions of $\hat{q}$ values may significantly vary when coancestry varies, when loci characterized by low and high allelic diversity are used, and when inbreeding is non-zero (Table 2).

Results from simulations for the 2, 6 and 10 allele cases differed markedly. Assignment error rate decreased (Figure 1(a–c)) or were approximately constant (Figure 1(d-e)) with increasing coancestry for the 2 allele case. The relationship was more complex for 6 and 10 allele cases, where error rates were greater at the high than at the low coancestry level (Figure 1). High variance in assignment accuracy was most notable in the 6 and 10 allele cases with increasing coancestry (Figure 1). Large variance in assignment error rates associated with high coancestry in simulated data sets can be explained by inter-family variance, as documented in the empirical case studies. In simulated data sets, individuals in a family were frequently all correctly or incorrectly classified (specific data not shown) as observed for empirical data sets (quantified using χ^2 tests).

Confamilial-based non-independence in assignment accuracy resulted in high inter-replicate variance in estimates of assignment accuracy, particularly for simulated data sets involving loci with higher allelic diversity (6 and 10 alleles per locus). Across simulated data sets, individuals (full-sibs) in each family were either largely correctly or mostly all misclassified (Figure 1). Results presented a trend toward bimodality in assignment accuracy (data not shown); one 'mode' representing low error rates whereas the other 'mode' depicted high error rates. These observations have important implications to empirical studies in natural populations. Mean assignment error rates described in Figure 1 for data sets with moderate or high allelic diversity (6 and 10 alleles per locus), and when source

Journal : GENE | Dispatch : 8-12-2005 | Pages : 12
CMS No. : DO00025365 | ☐ LE | ☐ TYPESET
MS Code : GENE150R2 | ☑ CP | ☑ DISK

populations are highly genetically differentiated (high F_{st} cases) should be interpreted with caution. Further, sampling of unknown individuals may lead to very different estimates of source population contributions if levels of coancestry are greater than zero.

Based on results from the simulation studies, estimates of accuracy of assignment described for empirical lake trout cases are readily explained. For the high F_{st} case (SMD and SLW) and for the low F_{st} case (SMD and SIW), allele distributions did not completely match because rare alleles were present. In simulated data sets, rare alleles may decrease accuracy of assignment (Cornuet et al., 1999), as the accuracy of allele frequency estimates is highly dependent on sample size (Guinand et al., 2004). In the low coancestry case, individuals possessing common alleles were sampled with higher probability. As a consequence, allelic distributions across families were more likely to overlap in situations of low coancestry and high F_{st}. Thus, assignment accuracy across families was not improved (Table 1). In cases of high coancestry and high F_{st}, individuals sampled had higher probabilities of sharing rare alleles and accuracy of assignment was greatly improved (Table 1).

Conclusion

Results of our empirical and simulated data sets have important implications for analyses of wild populations when biologists seek to correctly assign individuals to populations of origin. Assignment error is highly non-random when significant levels of coancestry existed among unknowns. Estimates of assignment accuracy made on the basis of resampling source populations will not likely be predictive of expected accuracy for unknown individuals characterized by non-zero coancestry. Gene correlations among individuals within source populations will upwardly bias estimates of assignment accuracy because the variance among groups (families or populations) is proportional to the coancestry of individuals within groups (i.e., gene correlation between individuals of the same breeding group; Cockerham, 1969, 1973). Genetic correlations among individuals will likely bias results unless assignment tests are accompanied by additional analyses that provide surrogate estimates of pedigree relationships (e.g., use of genetic markers to estimate coefficients of relationship such as Queller and Goodnight's r_{xy}; Queller and Goodnight, 1989).

Acknowledgements

We acknowledge U.S. Fish and Wildlife Service for providing samples and crosses of the lake trout strains used in this study. Constructive comments by F. Bonhomme and J.-D. Durand improved the quality of results presented here. This study was supported by the Great Lakes Fishery Trust, the Great Lakes Protection Fund, and by Partnership for Ecosystem Research and Management (PERM) between the Michigan Department of Natural Resources and Michigan State University, and by a grant of the Université des Sciences et Techniques du Languedoc to B.G.

References

Angers, B. & L. Bernatchez, 1996. Usefulness of heterologous microsatellites obtained from brood charr, *Salvelinus fontinalis* Mitchill, in other *Salvelinus* species. Mol. Ecol. 5: 317–319.

Arnold, K.E., 2000. Kin recognition in rainbow fish (*Melanotaenia eachamensis*): sex, sibs and shoalings. Behav. Ecol. Sociobiol. 48: 385–391.

Bekkevold, D., M.M. Hansen & V. Loeschke, 2002. Male reproductive competition in spawning aggregations of cod (*Gadus morhua*, L.). Mol. Ecol. 11: 91–102.

Belkhir, K., P. Borsa, L. Chikhi, N. Raufaste & F. Bonhomme, 2001. Genetix 4.02, logiciel sous Windows™ pour la génétique des populations. UMR CNRS 5171 Génome, Populations et Interactions, Université des Sciences et Techniques du Languedoc, Montpellier, France.

Bernatchez, L. & P. Duchesne, 2000. Individual-based genotype analysis in studies of parentage and population assignment: how many loci, how many alleles? Can. J. Fish. Aquat. Sci. 57: 1–12.

Berry, O., M.D. Tocher & S.D. Sarre, 2004. Can assignment tests measure dispersal? Mol. Ecol. 13: 551–561.

Castric, V. & L. Bernatchez, 2004. Individual assignment test reveals differential restriction to dispersal between two salmonids despite no increase of genetic differences with distance. Mol. Ecol. 13: 1299–1312.

Castric, V., L. Bernatchez, K. Belkhir & F. Bonhomme, 2002. Heterozygote deficiences in small lacustrine populations of brook charr *Salvelinus fontinalis* Mitchill (Pisces, Salmonidae): a test of alternative hypotheses. Heredity 89: 26–35.

Chesser, R.K., 1991a. Gene diversity and female philopatry. Genetics 127: 437–447.

Chesser, R.K., 1991b. Influence of gene flow and breeding tactics on gene diversity within populations. Genetics 129: 573–583.

Cockerham, C.C., 1969. Variance of gene frequencies. Evolution 23: 72–84.

Cockerham, C.C., 1973. Analysis of gene frequencies. Genetics 74: 679–700.

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

Efron, B., 1983. Estimating the error rate of a prediction rule: improvement of cross-validation. J. Am. Stat. Assoc. 78: 316–330.

Falush, D., M. Stephens & J.K. Pritchard, 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164: 1567–1587.

Fontaine, P.-M. & J.J. Dodson, 1999. An analysis of the distribution of juvenile Atlantic salmon (*Salmo salar*) in nature as a function of relatedness using microsatellite. Mol. Ecol. 8: 189–198.

Fraser, D.J., P. Duchesne & L. Bernatchez, 2005. Migratory charr schools exhibit population and kin associations beyond juvenile stages. Mol. Ecol. 14: 3133–3146.

Gerlach, G., U. Schardt, R. Eckmann & A. Meyer, 2001. Kin-structured subpopulations in Eurasian perch (*Perca fluviatilis* L.). Heredity 86: 213–221.

Guo, S.W. & E.A. Thompson, 1992. Performing the exact test of Hardy–Weinberg proportion for multiple alleles. Biometrics 48: 361–372.

Guinand, B., K.T. Scribner, A. Topchy, K.S. Page, W. Punch & M.K. Burnham-Curtis, 2004. Sampling issues affecting accuracy of likelihood-based classification using genetical data. Environ. Biol. Fish. 69: 245–249.

Hansen, M.M. & L.F. Jensen, 2005. Sibship within samples of brown trout (*Salmo trutta*) and implications for supportive breeding. Conserv. Genet. 6: 297–305.

Hansen, M.M., E.E. Nielsen & K.L.D. Mensberg, 1997. The problem of sampling families rather than populations: relatedness among individuals in samples of juvenile brown trout *Salmo trutta* L. Mol. Ecol. 6: 469–474.

Hansen, M.M., E. Kenchington & E.E. Nielsen, 2001a. Assigning individual fish to populations using microsatellite DNA markers. Fish Fish. 2: 93–112.

Hansen, M.M., D.E. Ruzzante, E.E. Nielsen & K.L.D. Mensberg, 2001b. Brown trout (*Salmo trutta*) stocking impact assessment using microsatellite DNA markers. Ecol. Appl. 11: 148–160.

Hedgecock, D., 1994. Does variance in reproductive success limit effective population sizes of marine organisms? pp. 122–134 in Genetics and Evolution of Marine Organisms, edited by A.R. Beaumont. Chapman & Hall, London.

King, T.L., S.T. Kalinowski, W.B. Schill, A.P. Spidle & B. Lubinski, 2001. Population structure of Atlantic salmon (*Salmo salar* L.): a range-wide perspective from microsatellite DNA variation. Mol. Ecol. 10: 807–821.

Kolm, N., E.A. Hoffman, J. Olsson, A. Berglund & A.G. Jones, 2005. Group stability and homing behavior but no kin group structures in a coral reef fish. Behav. Ecol. 16: 521–527.

Koskinen, M.T., J. Piironen & C.R. Primmer, 2001. Interpopulation genetic divergence in European grayling (*Thymallus thymallus*, Salmonidae) at a microgeographic scale: implications for conservation. Conserv. Genet. 2: 133–143.

Krause, J., R.K. Butlin, N. Peukhuri & V.L. Pritchard, 2001. The social organization of fish shoals: a test of the predictive power of laboratory experiments for the field. Biol. Rev. Camb. Phil. Soc. 75: 477–501.

Luikart, G. & P.R. England, 1999. Statistical analysis of microsatellite DNA data. Trends Ecol. Evol. 14: 253–256.

Manel, S., O.E. Gaggiotti & R.S. Waples, 2005. Assignment methods: matching biological questions with appropriate techniques. Trends Ecol. Evol. 20: 136–142.

Martinez, J.L., J. Dumas, E. Beall & E. Garcia-Vazquez, 2001. Assessing introgression of foreign strains in wild Atlantic salmon populations: variation in microsatellites assessed in historic scale collections. Freshwat. Biol. 46: 835–844.

Nielsen, E.E., M.M. Hansen, C. Schmidt, D. Meldrup & P. Grønkjaer, 2001. Population of origin of Atlantic cod. Nature 413: 272.

Nielsen, E.E., M.M. Hansen, D.E. Ruzzante, D. Meldrup & P. Grønkjaer, 2003. Evidence of a hybrid-zone in Atlantic cod (*Gadus morhua*) in the Baltic and the Danish Belt Sea revealed by individual admixture analysis. Mol. Ecol. 12: 1497–1508.

Olsen, J.B., P. Bentzen & J.E. Seeb, 1998. Characterization of seven microsatellite loci derived from pink salmon. Mol. Ecol. 7: 1087–1089.

O'Reilly, P.T., L.C. Hamilton, S.K. McConnel & J.W. Wright, 1996. Rapid analysis of genetic variation in Atlantic salmon (*Salmo salar*) by PCR multiplexing of dinucleotide and tetranucleotide microsatellite. Can. J. Fish. Aquat. Sci. 53: 2292–2298.

Paetkau, D., W. Calvert, I. Stirling & C. Strobeck, 1995. Microsatellite analysis of population structure in Canadian polar bears. Mol. Ecol. 4: 347–354.

Paetkau, D., R. Slade, M. Burden & A. Estoup, 2004. Genetic assignment methods for the direct, real-time estimation of migration rate: a simulation-based exploration of accuracy and power. Mol. Ecol. 13: 55–65.

Page, K.S., 2001. Genetic diversity and interrelationships of wild and hatchery lake trout in the upper Great Lakes: inferences for broodstock management and development of restoration strategies. M.S. Thesis, 121 pp. Department of Fisheries and Wildlife, Michigan State University, East Lansing.

Page, K.S., K.T. Scribner, K.R. Bennett & L.M. Garzel, 2003. Genetic assessment of strain-specific sources of lake trout recruitment in the Great Lakes. Trans. Am. Fish. Soc. 132: 877–894.

Planes, S. & P. Lenfant, 2002. Temporal change in the genetic structure between and within cohorts of a marine fish, *Diplodus sargus*, induced by a large variance in individual reproductive success. Mol. Ecol. 11: 1515–1524.

Planes, S., G. Lecaillon, P. Lenfant & M. Meekan, 2002. Genetic and demographic variation in new recruits of *Naso unicornis*. J. Fish Biol. 61: 1033–1049.

Pouyaud, L., E. Desmarais, A. Chenuil, J.-F. Agnèse & F. Bonhomme, 1999. Kin cohesiveness and possible inbreeding in the mouthbrooding tilapia *Sarotherodon melanotheron* (Pisces Cichlidae). Mol. Ecol. 8: 803–812.

Journal : GENE | Dispatch : 8-12-2005 | Pages : 12
CMS No. : DO00025365 | ☐ LE | ☐ TYPESET
MS Code : GENE150R2 | ☑ CP | ☑ DISK

Primmer, C.R., M.T. Koskinen & J. Piironen, 2000. The one that did not get away: individual assignment using microsatellite data detects a case of fishing competition fraud. Proc. R. Soc. Biol. Sci. B267: 1699–1704.

Queller, D.C. & K.F. Goodnight, 1989. Estimating relatedness using genetic markers. Evolution 43: 258–275.

Raymond, M. & F. Rousset, 1995. Genepop (Version 1.2): population genetics software for exact tests and ecumenicism. J. Hered. 86: 248–249.

Rice, W.R., 1989. Analysing tables of statistical tests. Evolution 43: 223–225.

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

Ross, K.G., 2001. Molecular ecology of social behaviour: analyses of breeding systems and genetic structure. Mol. Ecol. 10: 265–284.

Russell, S.T., J.L. Kelley, J.A. Graves & A.E. Magurran, 2004. Kin structure and shoal composition dynamics in the guppy (*Poecilia reticulata*). Oïkos 106: 520–526.

Ruzzante, D.E., M.M. Hansen & D. Meldrup, 2001. Distribution of individual inbreeding coefficients, relatedness and influence of stocking on native anadromous brown trout (*Salmo trutta*) population structure. Mol. Ecol. 10: 2107–2128.

Ryman, N., F.W. Allendorf & G. Ståhl, 1979. Reproductive isolation with little genetic divergence in sympatric populations of brown trout (*Salmo trutta*). Genetics 92: 247–262.

Storz, J.F., 1999. Genetic consequences of mammalian social structure. J. Mammal. 80: 553–569.

Sugg, D.W., R.K. Chesser, F.S. Dobson & J.L. Hoogland, 1996. Population genetics meets behavioral ecology. Trends Ecol. Evol. 11: 338–342.

Taylor, E.B., A. Kuiper, P.M. Troffe, D.J. Hoysak & S. Pollard, 2000. Variation in developmental biology and microsatellite DNA in reproductive ecotypes of kokanee, *Oncorhynchus nerka*: implications for declining populations in a large British Columbia lake. Conserv. Genet. 1: 213–249.

Taylor, E.B., Z. Redenbach, A.B. Costello, S.J. Pollard & C.J. Pacas, 2001. Nested analysis of genetic variation in northwestern North American char, Dolly Varden (*Salvelinus malma*) and bull trout (*S. confluentus*). Can. J. Fish. Aquat. Sci. 58: 406–420.

Waser, P.M. & C. Strobeck, 1998. Genetic signatures of interpopulation dispersal. Trends Ecol. Evol. 13: 43–44.

Weir, B.S., 1979. Inferences about linkage disequilibrium. Biometrics 25: 235–254.

Weir, B.S. & C.C. Cockerham, 1984. Estimating *F*-statistics for the analysis of population structure. Evolution 38: 1358–1370.

Wright, S., 1965. The interpretation of population structure by *F*-statistics with special regards to system of mating. Evolution 19: 395–420.

Wright, S., 1969. Evolution and the Genetics of Populations. Vol. 2: the Theory of Gene Frequencies . Chicago University Press, Chicago.

Journal : GENE | Dispatch : 8-12-2005 | Pages : 12
CMS No. : DO00025365 | ☐ LE | ☐ TYPESET
MS Code : GENE150R2 | ☑ CP | ☑ DISK