MOLECULAR ECOLOGY

Molecular Ecology (2011) **20**, 1976–1987 doi: 10.1111/j.1365-294X.2011.05069.x

Historical introgression and the role of selective vs. neutral processes in structuring nuclear genetic variation (AFLP) in a circumpolar marine fish, the capelin (*Mallotus villosus*)

GABRIEL J. COLBECK,* JULIE TURGEON,* PASCAL SIROIS† and JULIAN J. DODSON*

*Département de biologie, Pavillon Vachon, 1045 avenue de la Médecine, Université Laval, Québec, QC G1V 0A6, Canada, †Département des sciences fondamentales, 555 boulevard de l'Université, Université de Québec à Chicoutimi, Chicoutimi, QC G7H 2B1, Canada

Abstract

The capelin (*Mallotus villosus*) is a widespread marine fish species for which previous work has identified geographically distinct mtDNA clades, the frontiers of which are well within adult and larval dispersal capabilities. Here, we use AFLPs to test for the presence of nuclear gene flow among clades. In addition, we evaluate genetic structuring within one clade, the Northwest Atlantic (NWA). We found that each of the mtDNA clades corresponds with a unique nuclear DNA genetic cluster. Within the NWA clade, we detected individuals with small but significant amounts of genetic ancestry from other clades, likely due to historical introgression. Further support for historical introgression comes from analyses of variance in locus-specific differentiation, which support introgression between some clades and divergence without gene flow between others. Within the NWA, we identified two genetic clusters that correspond to sites in geographically adjacent areas. However, these clusters differ primarily at 'outlier' loci, and a genetic subdivision ($K = 2$) was not supported by genetic clustering programs using neutral loci. Significant neutral F_{ST} differentiation was found only between sites that otherwise differed at outlier loci. Thus, these populations may be in the initial stages of 'isolation by adaptation'. These results suggest strong between-clade reproductive isolation despite opportunities for gene flow and support the hypothesis that selection can contribute to divergence in otherwise 'open' systems.

Keywords: genome scan, glacial cycles, introgression, isolation by adaptation, marine, speciation

Received 6 October 2010; revision received 29 January 2011; accepted 8 February 2011

Introduction

The role of glacial cycles during the Quaternary/Pleistocene in shaping patterns of biodiversity remains a topic of considerable interest and debate. A wealth of studies have revealed distinct lineages with long independent trajectories (geographical and demographical – Avise *et al.* 1998), but the postglacial fates of those lineages can vary greatly. Considering the secondary contact that can occur during glacial cycles, opportunities for once divergent gene pools to merge should be frequent (Futuyma 1987; Jansson & Dynesius 2002). Thus, lineages may come back into secondary contact and experience gene flow (e.g. Yang & Kenagy 2009; Murtskhvaladze *et al.* 2010; Spinks *et al.* 2010) or they may remain independent (e.g. Carstens & Knowles 2007; Goncalves *et al.* 2009; Griffiths *et al.* 2010).

In addition, independent lineages can themselves experience sub-structuring in the absence of obvious physical barriers to gene flow because of the effects of distance and/or selection. Molecular genome-scan methods that allow the identification of loci that may be under the effects of divergent selection (those that differ

Correspondence: Gabriel J. Colbeck, Fax: 418 656 2043; E-mail: gabriel.colbeck.1@ulaval.ca

among populations more than one would expect under neutral models of divergence) provide the opportunity to discriminate between neutral and selective processes in shaping genetic structure (e.g. Gaggiotti *et al.* 2009; Bradbury *et al.* 2010). The question of limited gene flow may be particularly relevant for 'open' marine systems where larval and adult dispersal capabilities can be quite large, but ecological variation may be pronounced (reviewed in Cowen & Sponaugle 2009; Nielsen *et al.* 2009a). For example, Bradbury *et al.* (2010) found that Atlantic cod populations experience divergent selection associated with temperature across areas well within the range of adult dispersal.

The capelin (*Mallotus villosus*) is a circumpolar marine fish for which previous work has identified strong mtDNA divergence (dating to approximately 2 Ma) between clades with geographically distinct ranges (Dodson *et al.* 2007). The reciprocal monophyly of these lineages is surprising, given the short geographical distances between areas occupied by distinct clades (i.e. from Baffin Island to Hudson Bay) and the immense larval and adult dispersal capabilities of capelin (Behrens *et al.* 2006; Praebel *et al.* 2008). Nonetheless, gene flow may occur among clades if there is (i) limited gene flow along contact zones and/or (ii) male-biased dispersal and gene flow. If these processes are sufficiently common, divergent nuclear gene pools could experience introgression. Previous work on capelin also identified a lack of mtDNA differentiation across populations within the Northwest Atlantic (NWA) (Dodson *et al.* 2007), a result that is unexpected given the strong ecological heterogeneity of NWA water masses generated by the Labrador current and the partial isolation of the Gulf of St Lawrence (Longhurst 1998).

Here, we examine patterns of nuclear genetic variation within and between distinct mtDNA clades of the capelin. If nuclear gene flow is minimal among clades, we expect to see strong nuclear divergence among clades that mirrors the mtDNA divergence. If, however, nuclear gene flow has occurred among clades, we expect to find individuals with genetic ancestry from other clades.

Within one clade, the NWA, nuclear genetic subdivision may be present despite the lack of mtDNA structure (e.g. Mila *et al.* 2010). Such structuring may be driven by either neutral and/or selective processes; thus, we look for loci potentially affected by selection (outlier loci). The presence of divergent outliers would suggest the action of divergent selection in different populations. If outlier loci and neutral loci reveal the same geographical patterns of structure, we have evidence for 'isolation by adaptation', whereby selection has led to barriers to neutral gene flow (Nosil *et al.* 2009). However, if neutral loci do not show the same patterns of differentiation as outlier loci, divergent selection may be relatively recent, such that neutral loci may not yet have reached migration–drift equilibrium.

Materials and methods

A total of 273 adults were collected from 13 sites representing the geographical distribution of three very distinct mtDNA clades (3.1–4.5% net sequence divergence at mtDNA CytB, Dodson *et al.* 2007); the NWA, Arctic (ARC) and Northeast Atlantic (NEA) clades (Fig. 1, Table 1).

Laboratory analysis

We extracted DNA from muscle tissue preserved in 95% ethanol with a standard proteinase K, phenol–chloroform procedures. We quantified DNA with a low-mass ladder (Invitrogen) on 2% EtBr-stained agarose gels, and all samples were diluted to a working concentration of approximately 200 ng/μL.

We amplified a 572-bp sequence of the mtDNA cytochrome b gene following Dodson *et al.* (2007). These sequences complemented those already available and have been deposited in GenBank (accession numbers HQ340243–HQ340283).

For the Amplified Fragment Length Polymorphism (AFLP) analysis, we used only well-preserved DNA samples with minimal 'smearing' following the high–molecular weight genomic band when examined on an agarose gel. We generated AFLP fragments using the

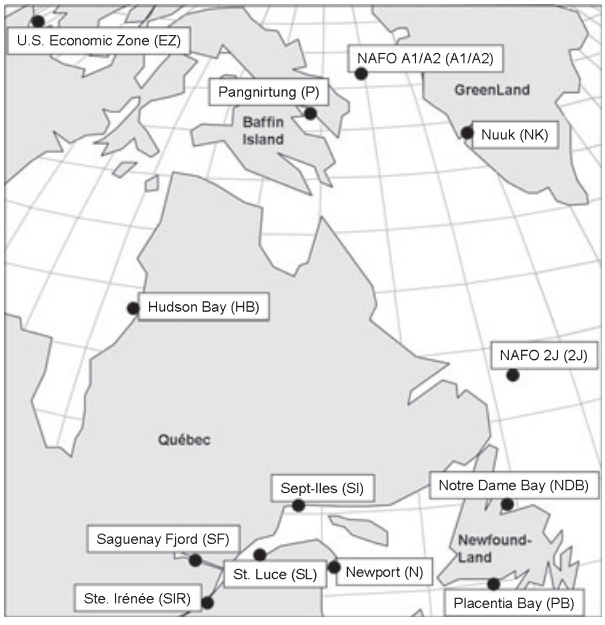

Fig. 1 Sample locations for capelin. See Table 1 for details on sampling sites.

Table 1 Sample locations and sizes for nuclear (AFLP) and mtDNA genetic characterization

Sample site	Latitude (°N)	Longitude (°W)	*N* (AFLP)	*N* (mtDNA)
Northwest Atlantic				
Newport (N)	48.28	64.16	29	3
Hudson Bay (HB)	56.16	76.47	9	
NAFO 2J (2J)	53.11	54.02	9	
Notre Dame Bay (NDB)	49.32	55.19	15	5
Placentia Bay (PB)	47.23	54.19	20	5
Saguenay Fjord (SF)	48.26	70.58	36	12
Sept-Iles (SI)	50.12	66.17	29	2
St Irénée (SIR)	47.33	70.11	29	3
St Luce (SL)	48.26	68.36	29	3
Arctic				
US Economic Zone (UE)	71.05	144.02	12	
Pangnirtung (P)	66.04	65.54	9	6
Northeast Atlantic				
NAFO 1B (A1/A2)	68.21	59.51	23	
Nuuk (NK)	64.08	52.06	24	2

Samples codes as per Fig. 1. DNA from five sampling sites (Hudson Bay, NAFO 2J, US economic zone, NAFO 1A and Nuuk, West Greenland) was obtained from an earlier sampling of capelin (see Dodson *et al.* 2007 for sampling details).

restriction enzymes *Eco*RI and *Mse*I (New England Biolabs, Ipswich, MA, USA) according to the protocol of Vos *et al.* (1995), with the following modifications. We conducted the PCRs using AFLP Core Mix (Applied Biosystems). We used an annealing temperature of 54 °C for the preselective PCR and diluted this product 1:10. We added 1 μL of this dilution to the final selective PCR, which consisted of 10 initial 'step-down' cycles beginning at 62 °C and ending at 52 °C, which were then followed by 25 cycles at 52 °C. Selective PCR were conducted with six primer pairs involving a fluorescently labelled *Eco*RI primer (Applied Biosystems) and an unlabelled *Mse*I primer (Sigma): ECO ACG + MSE CAG, ECO ACC + MSE CAG, ECO ATG + MSE CAG, ECO ACG + MSE CTC, ECO ACC + MSE CTC, ECO ACT + MSE CTC. We ran products from this selective amplification on an ABI 3100 capillary sequencer with LIZ size standard (Applied Biosystems) and collected the digital gel data using ABI Prism GeneMapper analysis software (version 3.75). Each lane file was analysed for the presence and absence of AFLP products with GeneMapper using a minimum relative fluorescent units (RFU) scoring threshold of 100 and verified by eye. We scored only those fragments between 50 and 400 bp, and we considered a locus to be scorable if the difference between presence and absence was discrete and the locus could be scored unambiguously. Per locus repeatability was calculated as the per cent of samples whose score (1 or 0) was identical between runs (for a total of 24 individuals re-run for the entire procedure). We discarded loci that were <90% repeatable. The final error rate for all retained loci was 3.2%. To control for the possibility of contamination, we included one negative control (a restriction/ligation containing no DNA, run through the entire procedure) in each plate of 96 individuals. The negative control never produced any amplified product.

Genetic analysis

mtDNA. Our objective in sequencing mtDNA was to determine the clade affiliation of samples new to this analysis. Given the very clear clade membership of any haplotype, we simply added new sequences to the original data set of Dodson *et al.* (2007) and performed the same phylogenetic analyses. Most individuals from the US Economic Zone, Nuuk and NAFO 1B were sequenced previously. Two individuals from Nuuk were sequenced for this study only because they had ARC instead of the expected NEA nuclear profiles (see Results).

AFLP. Detection of outlier loci. In an attempt to tease apart population genetic patterns that are attributable to selective vs. neutral processes, we first sought to identify loci that are potentially under selective pressure using the Bayesian approach implemented by BAYESCAN 1.0 (Foll & Gaggiotti 2008). This methodology estimates the probability that a locus is experiencing selection by directly comparing the posterior probabilities of two models: one that includes the effects of selection and one that does not (Foll & Gaggiotti 2008). In addition, BAYESCAN uses estimates of local F_{ST} (as opposed to pairwise F_{ST}) and is robust to complex demographic models. To minimize the detection of false positives (see Foll & Gaggiotti 2008; Excoffier *et al.* 2009), we inferred

the presence of outliers using the 10 sample sites within the NWA (for which mtDNA indicated no substructure), and only interpreted outliers identified at the Bayes factor = 99.0 threshold. Long periods of isolation followed by rapid population expansion precluded us from conducting an outlier analysis between mtDNA clades, a situation that can drastically elevate the appearance of false positives (Foll & Gaggiotti 2008). Simulation studies indicate the presence of little to no false positives among outliers detected at the Bayes factor = 99 threshold and suggest that BAYESCAN is much better at avoiding false positives than other methodologies (Perez-Figueroa *et al.* 2010). Nonetheless, for the sake of comparison, we also used FDIST2 (Beaumont & Nichols 1996) to identify outliers.

Because bands at the two most divergent outlier loci identified within the NWA appeared in other clades, as well as in isolated populations (e.g. the Saguenay and Newfoundland), we sought to determine whether these bands were homologous (i.e. were not size homoplasious). We conducted a second round of four selective amplifications (+4 reactions), each with the original selective MSE primer plus one additional base pair (MSE CAG + A, CAG + C, CAG + G and CAG + T) along with the original ECO + 3 primer (O'Hanlon & Peakall 2000). The logic is that the next base pair in the sequence of the original selectively produced fragment will be one of the four bases and thus should be reproduced in this second round by only one of the four +4 reactions. If different +4 reactions produce a fragment in different populations (or clades), then we have evidence for size homoplasy. This test was particularly important for our between-clade analysis, as the probability of size homoplasy increases with taxonomic distance (Bonin *et al.* 2007), and the distinct capelin clades have been isolated for ~2 Myr. We conducted the +4 reactions on 10 individuals known to have a zero at the outlier loci, and five individuals from each clade known to have a band at the outlier loci, for a total sample size of 25.

We next tested for associations between outlier loci and environmental variables with a method analogous to that of GESTE (Foll & Gaggiotti 2006). In GESTE, if statistical associations between site-specific F_{ST} values and environmental variables are found in analyses with outlier loci, but not without outlier loci, then those outlier loci are assumed to be specifically involved with selection related to the environmental variables. Because GESTE does not accommodate dominant (AFLP) data, we sought to use an analogous method. First, we used ABC4F (Foll *et al.* 2008) to calculate site specific F_{ST} values. Next, we conducted principal components analysis on environmental variables as well as latitude and longitude for sampling sites (because adults were captured while spawning) within the NWA. The environmental variables (Table S3A, Supporting information) included salinity and temperature at the beginning and end of known spawning activity (May and August) at the water surface and at a depth of 50 m (as Capelin can spawn on the beach as well as at depth – e.g. Carscadden *et al.* 1989) and were obtained from the Canadian Department of Fisheries and Oceans: Database of Salinity and Temperature Observations for the NWA (http://www2.mar.dfo-mpo.gc.ca/science/ocean/tsdata.html). We next conducted a regression on local F_{ST} and environmental Principal Component (PC) scores.

We also used the more traditional method of a Mantel test as implemented in GENALEX 6.2. We used a matrix of pairwise F_{ST} values from AFLPSURV and a matrix of pairwise differences in environmental PC scores as obtained earlier.

Genetic structure among and within clades. To detect population genetic structure, we utilized both genetic clustering methods and indices of differentiation (i.e. pairwise F_{ST}). To assess the effects of selection on population genetic structure, we performed all within-clade analyses both with and without outlier loci. For clustering methods, we first used a principal components analysis as implemented in GENALEX 6.0 (Peakall & Smouse 2006) to visualize structure among and within clades. Next, we used the program STRUCTURE 2.3 (Pritchard *et al.* 2000). Given the nature of our data set and following the recommendations of Hubisz *et al.* (2009), we chose to use both the original (no LOCPRIOR, Pritchard *et al.* 2000) as well as the new (with LOCPRIOR, Hubisz *et al.* 2009) models implemented by STRUCTURE 2.3. On the one hand, given the level of divergence among clades, confirmed by F_{ST} values (see Results), we have highly informative data, for which the original model is more appropriate. On the other hand, the original model of STRUCTURE can have difficulty finding a genetic signal from 'small' sample sizes, and some of our sites had as few as nine individuals. We used a burn in of 10 000, chain length of 100 000 and the 'admixture' model for all runs (increasing burn in and chain length up to 100 000 and 1 000 000, respectively, did not change estimation of K or geographical distribution of genetic clusters). To determine the true value of K, we used two criteria that provided similar estimates (Pritchard *et al.* 2000 and Evanno *et al.* 2005; Hubisz *et al.* 2009) and report the results as mean likelihood ± $\sigma^2/2$ for three Ks: the K preceding the most likely ($K - 1$), the most likely K, and the K following the most likely ($K + 1$). Given some differences between results with and without LOCPRIOR, we also used FLOCK v. 1.0 (Duchesne & Turgeon 2009), an algorithm using maximum-likelihood iterative allocation to estimate the K (the number of genetic clusters). We evaluated the most likely value of

K via an log-likelihood difference (LLOD) plateau analysis. For a given *K*, FLOCK performs 50 runs and provides the LLOD of each of the 50 runs. When two or more runs have the same LLOD, they form a 'plateau' and indicate runs in which individuals are identically allocated into clusters. Simulations indicate a plateau of at least six is indicative of real structure (Duchesne & Turgeon, unpublished data), so the largest value of *K* with a LLOD plateau of at least six is a highly probable value of *K*. The use of FLOCK along with STRUCTURE provided a measure of confidence in estimating *K*.

We also estimated genetic structure by evaluating nuclear differentiation between and within mtDNA clades with F_{ST} analyses implemented by AFLP-SURV (Vekemans 2002). For the AFLP-SURV analysis, we used the Bayesian method of generating band frequencies (Zhivotovsky 1999) with a nonuniform prior distribution. We determined the significance of pairwise differentiation with Bonferroni correction for multiple comparisons.

Tests for historical introgression. To specifically test for the presence of individuals with mixed ancestry, we evaluated whether some individuals could qualify as F1 or backcross hybrids between mtDNA clades. To do so, we used the F1/F2 allocation procedure of AFLPOP (Duchesne & Bernatchez 2002; e.g. den Hartog *et al.* 2010; McKinnon *et al.* 2010). We first defined reference parental populations as the 30 most likely individuals from each clade (identified by FLOCK). We then allocated all individuals among these parental references as well as the F1, parental × F1 and F2 classes. Then, we compared allocations rates with expected error rates using the simulation option of AFLPOP; multilocus genotypes of each class (i.e. parental, F1, F2, etc.) are generated and simulated individuals are re-allocated to these classes. The proportion of incorrect allocations (for example, a simulated individual from a parental population who is allocated to the F1 population) is the error rate and indicates a lack of statistical power to discriminate between classes. However, to the extent that the empirical rate of assignments exceeds the error rate, we have strong evidence for 'hybrid' assignment. Because the number of individuals in the NEA and ARC clades was considerably reduced compared to the NWA, we first performed hybrid assignments with all data and then separately with equal sample sizes for each clade (Table S1, Supporting information). Finally, we compared the coefficients of ancestry (from STRUCTURE) between individuals with putative hybrid ancestry and those without putative hybrid ancestry with an ANOVA. We determined significant differences between groups (clade 1, hybrids and clade 2) with Tukey *post hoc* comparisons.

The assignment of real individuals to simulated hybrid populations (as per the above-mentioned analysis) could indicate that gene flow has occurred between clades or alternatively could arise because these individuals retained ancestral polymorphisms. Wakeley & Hey (1997) and Hey & Nielsen (2004) have demonstrated that it is possible to distinguish between these scenarios based on the statistical properties of multiple, independently segregating loci. At the most basic level, introgression should result in some loci being relatively undifferentiated, while others remain highly divergent (resulting in a large variance in differentiation across loci), whereas incomplete lineage sorting should result in a more uniform distribution of differentiation across loci (Hey 2006). Unfortunately, because of the unknown mutation mechanisms and dominant nature of AFLPs, it is not possible to use the methodology of Hey & Nielsen (2004), so we attempted a more qualitative approach. We first used BAYESCAN to estimate locus-specific F_{ST} differentiation between each pair of clades. We then generated frequency distributions and estimated standard deviations of locus-specific differentiation indices for each pair of clades. Thus, pairs of clades that have experienced introgression should have a higher variance in locus-specific differentiation than those that experienced divergence without gene flow. To avoid potential problems caused by differences in sample sizes between clades, we perform this analysis with all sample sites as well as with one sample site per clade.

Results

Genetic structure among clades

All new haplotypes generated in this study fell perfectly within the clade system described by Dodson *et al.* (2007) with the exception of two individuals sampled in Nuuk, West Greenland (where all other individuals had NEA haplotypes) who had typical ARC clade haplotypes. All of the new sequences from Baffin Island ($n = 6$) fell within the ARC clade, and all new sequences from the St Lawrence Estuary/Gulf and Newfoundland ($n = 33$) fell within the NWA clade.

Six AFLP primer combinations yielded a total of 409 bands between 50 and 400 bp of which 319 varied discretely between presence and absence and were thus deemed objectively scorable. Two hundred and thirty-one of those loci had repeatabilities of 90.0% or higher (i.e. produced different results in a maximum of two out of the 24 repeated individuals), 214 of which were variable between clades, and 182 of which were variable within the NWA. Some 134 loci (58%) were polymorphic between 5% and 95% between clades, and 146 (63%) were polymorphic within the NWA.

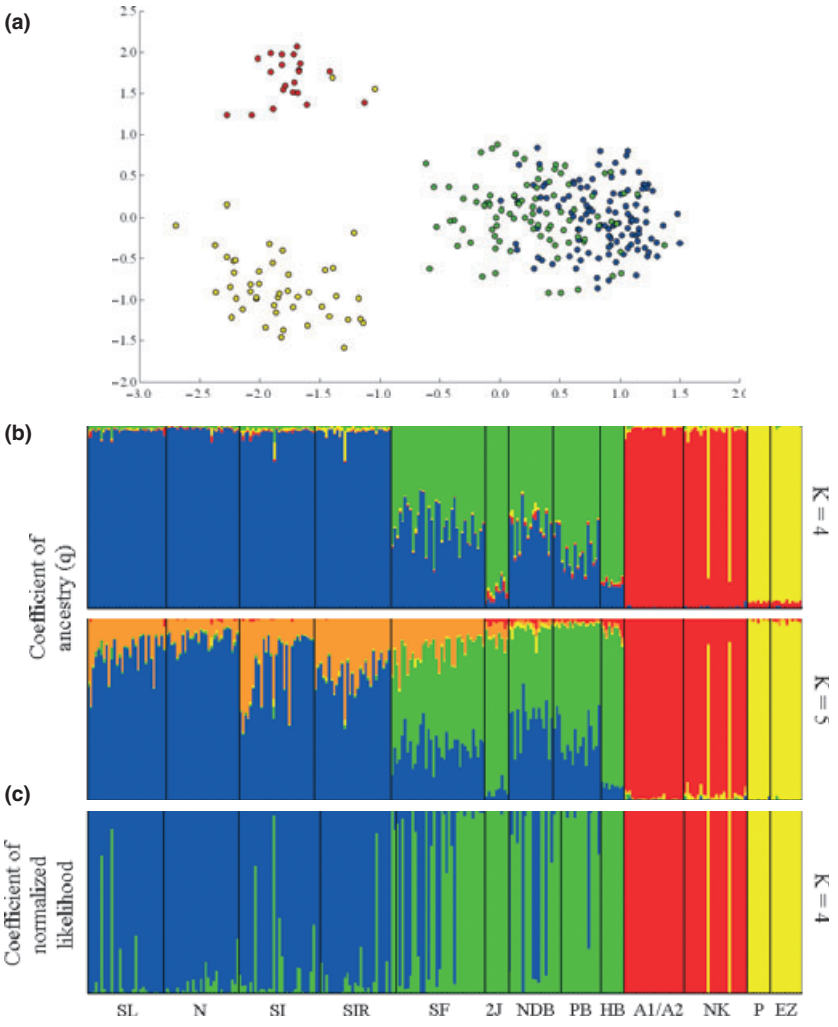

Fig. 2 Genetic clusters as ascertained with (a) principal components analysis, red = Northeast Atlantic, yellow = Arctic, blue = St Lawrence Estuary/Gulf, green = Saguenay, Newfoundland, 2J and Hudson Bay; (b) STRUCTURE using site as LOCPRIOR, $K = 4$ and 5; and (c) FLOCK, $K = 4$.

Considering all analyses, there appear to be four AFLP clusters, two that correspond to two of the mtDNA clades and two others found within the NWA (see below). In the principal components analysis, individuals from each of the clades form distinct clusters, while two overlapping clusters are apparent within the NWA (Fig. 2a). The most likely number of genetic groups in the STRUCTURE analyses were $K = 3$ with no LOCPRIOR (Figs. S1 & S3, Supporting information; $K_2 = -5242.5 \pm 53.7$, $K_3 = -5090 \pm 112.7$, $K_4 = -4998 \pm 175.2$) and $K = 5$ when site is used as a LOCPRIOR (Fig. 2b, $K_4 = -5074 \pm 66.9$, $K_5 = -4720 \pm 189.9$, $K_6 = -4689.8 \pm 303.7$). Without information on site, STRUCTURE failed to detect the divergence between the ARC and NEA clades at $K = 3$ (see Fig. S1, Supporting information), even though they are highly divergent (Table 2, Fig. 2a,c, see Supporting information). When using site as LOCPRIOR, at $K = 5$, two of the genetic clusters correspond exactly with the ARC and NEA clades, while the mtDNA NWA clade comprises three subclusters, one of which is distributed across nearly all sites (Fig. 2b). The most likely number of genetic clusters in the FLOCK analysis is $K = 4$ (Fig. 2c, likelihood plateau length = 6), whereas Ks of five through eight yielded no likelihood plateaus. The geographical distribution of the genetic clusters in the FLOCK analysis perfectly corresponds to mtDNA clades (but with two genetic clusters in the NWA, as per Principal Components Analysis, Fig. 2a) and coincides quite well with the STRUCTURE analysis for

Table 2 Interclade pairwise AFLP F_{ST} comparisons

	NWA	NEA	ARC
NWA			
NEA	0.18		
ARC	0.29	0.24	

Values come from AFLP-SURV (all significant at $P < 0.001$).
ARC, Arctic; NEA, Northeast Atlantic; NWA, Northwest Atlantic.

Table 3 Assignment rates and error rates of AFLPOP hybrid allocations

	ARC (ref 1)		NWA (ref 2)		NEA (ref 1)		NWA (ref 2)	
	Actual (%)	Error (%)	Actual (%)	Error (%)	Actual (%)	Error (%)	Actual (%)	Error (%)
F1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
F1 × Ref1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
F1 × Ref2	0.0	0.0	0.0	3.8	0.0	0.0	0.0	2.2
F2	0.0	0.0	5.9	0.0	0.0	0.0	19.3	0.0

Percentage of real individuals (actual) and simulated individuals (error) assigned to simulated hybrid populations. The left and right panels consider simulated hybridization between the ARC and NWA, and NEA and NWA, respectively.
ARC, Arctic; NEA, Northeast Atlantic; NWA, Northwest Atlantic.

$K = 4$ using site as LOCPRIOR (Fig. 2b). In the analysis of pairwise F_{ST} among clades, each clade is highly and significantly differentiated from the others (Table 2).

Interestingly, the two individuals sampled in Nuuk with ARC mtDNA also have ARC AFLP profiles (Fig. 2a). In Fig. 2b ($K = 4$), these two individuals appear to be of mixed ancestry, with ~25% ancestry in the NEA. This, however, is purely an effect of the use of site as a LOCPRIOR; when site is not used as a LOCPRIOR, these individuals have pure ARC profiles (Fig. S1b,c, Supporting information). In addition, they have 'pure' ARC profiles in the FLOCK analysis (Fig. 2c).

Analyses of introgression

One means of assessing the existence and/or prevalence of gene flow between highly divergent groups is by simulating hybrid populations and then performing assignment tests among pure parental and simulated hybrid populations (e.g. Peccoud *et al.* 2009). Assignment of simulated interclade hybrids and analyses of variance in locus-specific differentiation both suggest historical introgression between NWA and each of NEA and ARC. For the between-clade hybrid simulations, low assignment error rates were observed, and the actual rate of assignment always far exceeded the error rate (Table 3), suggesting we had high power for the classification of hybrid individuals. We found no putative hybrids between the ARC and NEA. We found 8 (5.9%) putative hybrids between the ARC and NWA (all sampled in the NWA) and all assigned to the F2 population. We also found 26 (19.3%) putative hybrids between the NEA and the NWA (all sampled in the NWA) and all assigned to the F2 population (Table 3). Although we did not detect any individuals in the ARC or NEA with putative NWA hybrid ancestry, our limited sample sizes in these clades prevented us from discounting this possibility. The effect of a smaller sample size was seen when we reduced the sample size from the NWA to equal that of the ARC and our ability to detect hybrid individuals was reduced (Table S1, Supporting information).

The coefficients of ancestry (from STRUCTURE) for individuals with putative hybrid ancestry were clearly intermediate: they were significantly higher than 'pure' individuals from the NWA and significantly lower than 'pure' individuals from the other clades (Fig. 3a,b – all comparisons are significant at $P < 0.001$) Logically, these analyses confirmed that these individuals had a dominant NWA nuclear ancestry, as per clustering analyses.

F_{ST} indices of differentiation were more variable between clades for which historical hybridization was suspected. The standard deviations in locus-specific F_{ST} between the clades were as follows: NEA/NWA, $\sigma = 0.062$; ARC/NWA, $\sigma = 0.067$; ARC/NEA, $\sigma = 0.037$. Thus, the standard deviation of locus-specific differentiation between the ARC and NEA is nearly half that of the standard deviations between the NWA and ARC and NWA and NEA (Fig. 4). Results shown are based on one sample site per clade of approximately equal sample size and were qualitatively similar, no matter which sample sites were used.

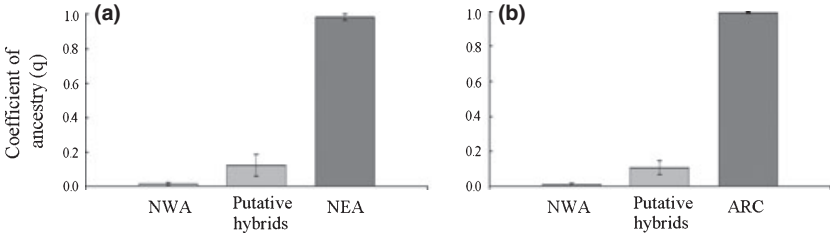

Fig. 3 Mean ± SD coefficients of ancestry from (a) Northeast Atlantic (NEA) and from (b) Arctic (ARC) for 'pure' Northwest Atlantic individuals, putative hybrids (identified by AFLPOP) and 'pure' (a) NEA and (b) ARC individuals.

Genetic structure within the NWA

We identified four outlier loci at Bayes' factor = 99.0 (CAG-ACG 06, F_{ST} = 0.31; CAG-ACG 68, F_{ST} = 0.18; CTC-ACC 52, F_{ST} = 0.15; CTC-ACG 07, F_{ST} = 0.13 – Fig. 5). The same four loci are identified at $P < 0.01$ with FDIST2 (Fig. S2, Supporting information). The addition of one base pair to the selective amplifications revealed that bands at the two most divergent outlier loci are homologous across regions and clades. For locus CAG-ACG 06, only the addition of a 'G' to the CAG primer reproduced the originally scored band, and for locus CAG-ACG 68, only the addition of a 'C' reproduced the originally scored band. PCR product was present with the addition of other base pairs, but it was irregularly shaped and had much lower intensity (RFU) than the originally scored band.

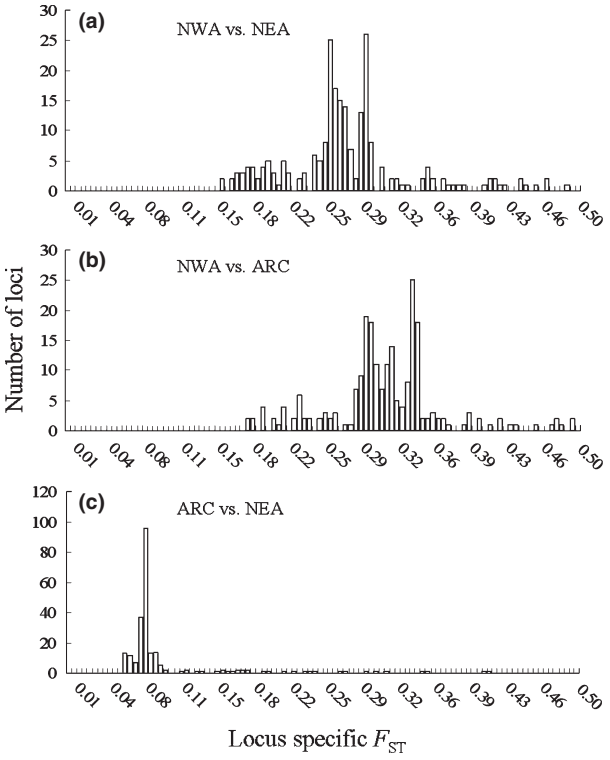

Fig. 4 Frequency distributions of locus-specific F_{ST} values for analyses comparing (a) Northwest Atlantic (NWA) to Northeast Atlantic (NEA) sites, (b) NWA to Arctic (ARC) sites and (c) NEA to ARC sites.

No significant associations were found between environmental variables and outlier loci (Table S3, Supporting information).

With all loci, clustering analyses revealed two genetic clusters within the NWA: one is found in sites associated with the St. Lawrence system (Ste. Irénée to Newport and Sept-Iles), while the other is found in sites from Newfoundland to Hudson Bay as well as in the Saguenay Fjord (Fig. 2b – K = 4, Fig. 2c).

Pairwise F_{ST} analyses within the NWA largely confirm the groupings found by STRUCTURE and FLOCK. After controlling for multiple comparisons, no significant differences exist between the Saguenay and Newfoundland to Hudson Bay, and no differences exist between Ste. Irénée, Ste. Luce, Sept-Iles and Newport, but significant differences exist between sites of these two groups (Table 4).

When outlier loci were removed, the most likely number of genetic groups in the STRUCTURE analyses was K = 1. In the FLOCK analysis, there was also no evidence

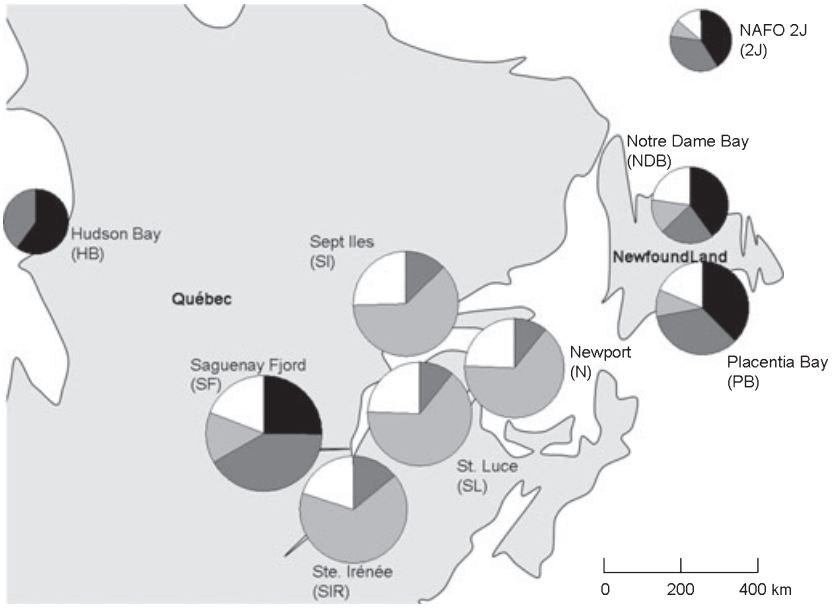

Fig. 5 Pie charts depicting the relative proportion of bands at the four outlier loci. Black = CAG-ACG 06, dark grey = CAG-ACG 68, light grey = CTC-ACC 52, white = CTC-ACG 07. The size of the circle is proportionate to the sample size (see Table 1).

Table 4 Pairwise AFLP F_{ST} comparisons within the Northwest Atlantic

	SIR	SL	SF	SI	N	PB	NDB	2J	HB
SIR									
SL	<0.001								
SF	0.047**	0.028							
SI	0.021	0.009	0.051**						
N	0.016	<0.001	0.048**	0.021					
PB	0.022	0.019	0.013	0.042**	0.031				
NDB	0.023	0.031	0.025	0.035	0.040**	<0.001			
2J	0.089**	0.062	0.056	0.086**	0.075**	0.035	0.037		
HB	0.082**	0.048	0.025	0.063**	0.049**	0.013	0.018	0.064	

Values come from AFLP-SURV.
**Significant at $P = 0.001$.

for two or more clusters. In the principal components analysis, the two overlapping groups within the NWA apparent in Fig. 2a merge into one cluster (Fig. S4, Supporting information). Thus, the presence of two genetic clusters in the NWA is driven by loci that are highly likely to be under divergent selection.

After removing outlier loci, there was no significant F_{ST} differentiation after controlling for multiple comparisons. We then pooled sites together that did not differ at $P = 0.05$ and re-conducted the analysis (Table S2, Supporting information). The new groups were the following: (i) Ste. Irene, Ste. Luce, Sept-Iles and Newport; (ii) Saguenay; and (iii) Notre Dame Bay, Placentia Bay, NAFO 2J and Hudson Bay. Significant F_{ST} differences remained between the Saguenay and sites associated with the St Lawrence Estuary/Gulf ($F_{ST} = 0.040$, $P < 0.001$), between the Saguenay and Newfoundland/Hudson Bay ($F_{ST} = 0.019$, $P = 0.004$) and between the St Lawrence Estuary/Gulf and Newfoundland/Hudson Bay ($F_{ST} = 0.032$, $P < 0.001$).

With all loci, the Saguenay Fjord is associated with sites in Hudson Bay and near Newfoundland that it is physically isolated from (Table 4, Fig. 2b,c). However, because the presence of two genetic clusters in the NWA is driven by loci that are likely to be under divergent selection, the association between the Saguenay Fjord and Hudson Bay/Newfoundland is because of outlier loci (and potentially similar selective pressures) as opposed to neutral patterns of gene flow.

Discussion

Glacial advances during periods of climate change almost certainly led to fragmentation of species' distributions, but as glaciers retreated, secondary contact may have diminished opportunities for speciation. For the capelin, strong evidence exists to support long periods of between-clade isolation during the Pleistocene (Dodson *et al.* 2007). However, these distinct clades now occupy adjacent areas (easily traversable by the high dispersal capabilities of capelin), and they have had the opportunity for secondary contact at least since the last glacial maximum. In this study, we find that nuclear gene flow is highly limited between capelin clades, suggesting that reproductive isolation is strong and is probably limited by factors other than distance. We identify adult dispersal (two individuals with ARC DNA profiles sampled near Nuuk, West Greenland, in the spring; thus, it is highly likely that they would have spawned in Greenland.). Nevertheless, no evidence of genetic admixture between these clades was found, and thus there is no evidence of gene flow.

We do, however, find individuals from the NWA who assign to simulated hybrid populations, suggesting that these individuals have genetic ancestry from other clades. Because these individuals assign to F2 populations (and not F1 populations), because their coefficients of ancestry from other clades are intermediate, but low (Fig. 3), and because we detect no dispersal events in or out of the NWA, the presence of genetic ancestry from other clades in the NWA is likely due to historical as opposed to recent introgression. Alternatively, these patterns could be produced by the retention of ancestral polymorphisms. Our results best support a model of historical introgression for two reasons. First, if the detection of genetic ancestry from other clades was simply a matter of lineage sorting, we should see the most genetic ancestry (and the most assignments to hybrid populations) in the case of populations with the shortest estimated time since divergence – the ARC and the NWA (Dodson *et al.* 2007). Instead, we see far and away more genetic ancestry and hybrid assignments between the most divergent (mtDNA) populations – the NEA and the NWA. Second, estimates of variance in locus-specific differentiation reveal much more variation between the NWA/ARC and NWA/NEA compared to the ARC/NEA (Fig. 4). These patterns support the scenario in which the ARC and NEA experienced diver-

gence without gene flow (i.e. most loci show the same level of differentiation), while the NWA experienced divergence and subsequent introgression with the other two clades (i.e. some loci show low differentiation, while others show high differentiation – Fig. 4). Importantly, these results are not an effect of the larger sample size in the NWA, as the larger variance is seen even when including only one sample site of equal sample size to the other clades.

Genetic differentiation owing to selection

Associations between ecological variables and outlier loci can sometimes be used to infer the nature of selection (reviewed in Nosil *et al.* 2009). For example, in highly vagile marine fishes, associations between outlier loci and ecological variables such as temperature and salinity (Gaggiotti *et al.* 2009; Nielsen *et al.* 2009b; Bradbury *et al.* 2010), depth (White *et al.* 2010) and even environmental pollution (Williams & Oleksiak 2008) have been used to infer adaptive divergence in otherwise open systems.

The outlier loci we have identified lead to the clustering of the Saguenay Fjord with sites from Newfoundland to Hudson Bay, and the clustering of sites within the St Lawrence Estuary/Gulf from Sept-Iles to Newport (Fig. 5). While neutral markers do not show the same level of differentiation across these same sample sites (i.e. multiple genetic clusters are not supported by STRUCTURE, FLOCK), we find significant F_{ST} differentiation without outlier loci only between population pairs that otherwise differ at outlier loci. These results do not fully support a scenario of 'isolation by adaptation', but they do suggest that the selective differences between sites associated with the Estuary/Gulf of St Lawrence and adjacent Newfoundland/Saguenay sites may be so recent to have only weakly impacted neutral markers. In other words, given the large effective population sizes that are likely to prevail for capelin, neutral loci may not yet have reached migration–drift equilibrium.

It is interesting that the divergent selection affecting loci in the St Lawrence Estuary/Gulf did not affect the same loci in the adjacent Saguenay Fjord. The Saguenay has long been considered an arctic refuge, possessing many arctic fishes and invertebrates not found in the St Lawrence Estuary (Judkins & Wright 1974), the persistence of which could be because of favourable hydrothermal conditions (Bosse *et al.* 1996). However, we cannot identify any association between outlier loci and temperature or salinity (Table S3, Supporting information). Thus, the selection affecting the presence of bands at the outlier loci in this system remains elusive.

Here, we have identified outlier loci that have diverged across adjacent sites and over distances well within the range of adult and/or larval dispersal, putatively because of the effects of divergent selection. Alternatively, it is possible that these loci are false positives and have diverged across sites because of the effects of drift in isolated populations. We feel that this is unlikely for several reasons. First, simulation studies suggest that very few, if any, outlier loci detected at the Bayes factor = 99 threshold are false positives (Perez-Figueroa *et al.* 2010). Second, we identify the exact same outlier loci at the most stringent statistical threshold of an alternative method, FDIST2 (Fig. S2, Supporting information). Third, frequencies of alleles at outlier loci are very similar in the Saguenay Fjord and Newfoundland. If our outliers are false positives, it requires the interpretation that these two populations are connected by neutral gene flow, a highly improbable scenario considering the sites are separated by the St Lawrence Estuary/Gulf, sites that have very different outlier allele frequencies (Fig. 5). And finally, mtDNA haplotype frequency analyses do not support the existence of isolated populations within the NWA (Dodson *et al.* 2007).

Conclusions

Despite the potential for historical climate fluctuations to lead to the merger of gene pools between once isolated populations, we found strong nuclear genetic isolation between geographically adjacent capelin clades. This result is surprising given the large dispersal capabilities of capelin and suggests that reproductive isolation is maintained by mechanisms other than pure distance. In addition, we found evidence to support historical introgression between certain clades, suggesting that gene flow may have occurred in the past but is now minimal. The mechanisms of reproductive isolation between capelin clades will require further investigation.

Within the NWA, we found differentiation across adjacent areas, which is likely due to divergent selection associated with as of yet unidentified conditions. With neutral loci, significant F_{ST} differentiation remained only between population pairs that otherwise differ at outlier loci. Thus, selection may be contributing to divergence across geographically adjacent areas that are in the initial stages of 'isolation by adaptation'.

Acknowledgements

This work was funded in part by a grant from the Canadian Healthy Oceans Network to JJD and PS. We thank the many researchers and wildlife biologists who assisted with sample collection in Baffin Island, Newfoundland and the St Lawrence Estuary/Gulf. Catherine Potvin provided help with laboratory

work, and Pierre Duchesne provided valuable assistance with data analysis. Audrey Bourret, Marc-Antoine Couillard, Marie-Claude Gagnon, David Paez, Genevieve Parent and two anonymous reviewers provided helpful discussion and comments on the manuscript. Data deposited at Dryad: doi:10.5061/dryad.8616.

References

Avise JC, Walker D, Johns GC (1998) Speciation durations and Pleistocene effects on vertebrate phylogeography. *Proceedings of the Royal Society of London Series B – Biological Sciences*, **265**, 1707–1712.

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society B – Biological Sciences*, **263**, 1619–1626.

Behrens JW, Praebel K, Steffensen JF (2006) Swimming energetics of the Barents Sea capelin (*Mallotus villosus*) during the spawning migration period. *Journal of Experimental Marine Biology and Ecology*, **331**, 208–216.

Bonin A, Ehrich D, Manel S (2007) Statistical analysis of amplified fragment length polymorphism data: a toolbox for molecular ecologists and evolutionists. *Molecular Ecology*, **16**, 3737–3758.

Bosse L, Saint-Marie B, Fournier J (1996) Les invertebres des fonds meubles et la biogeographie du fjord du Saguenay. *Rapport technique canadien des scences halieutiques et aquatiques*, **2132**, 45.

Bradbury IR, Hubert S, Higgins B *et al.* (2010) Parallel adaptive evolution of Atlantic Cod on both sides of the Atlantic Ocean in response to temperature. *Proceedings of the Royal Society B – Biological Sciences*, **277**, 3725–3734.

Carscadden JE, Frank KT, Miller DS (1989) Capelin (*Mallotus villosus*) spawning on the southeast shoal: influence of physical factors past and present. *Canadian Journal of Fisheries and Aquatic Sciences*, **46**, 1743–1754.

Carstens BC, Knowles LL (2007) Shifting distributions and speciation: species divergence during rapid climate change. *Molecular Ecology*, **16**, 619–627.

Cowen RK, Sponaugle S (2009) Larval dispersal and marine population connectivity. *Annual Review of Marine Science*, **1**, 443–466.

den Hartog PM, den Boer-Visser AM, ten Cate C (2010) Unidirectional hybridization and introgression in an avian contact zone: evidence from genetic markers, morphology and comparisons with laboratory raised F1 hybrids. *The Auk*, **127**, 605–616.

Dodson JJ, Tremblay S, Colombani F, Carscadden JE, Lecomte F (2007) Trans-Arctic dispersals and the evolution of a circumpolar marine fish species complex, the capelin (*Mallotus villosus*). *Molecular Ecology*, **16**, 5030–5043.

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.

Duchesne P, Turgeon J (2009) FLOCK: a method for quick mapping of admixture without source samples. *Molecular Ecology Resources*, **5**, 1333–1344.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, **14**, 2611–2620.

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity*, **103**, 285–298.

Excoffier L, Lischer HE (2010) Arlequin suite version 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, **10**, 564–567.

Foll M, Gaggiotti O (2006) Identifying the environmental factors that determine the genetic structure of populations. *Genetics*, **174**, 875–891.

Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, **180**, 977–993.

Foll M, Beaumont MA, Gaggiotti O (2008) An approximate Bayesian computation approach to overcome biases that arise when using amplified fragment length polymorphism markers to study populations structure. *Genetics*, **179**, 927–939.

Futuyma DJ (1987) On the role of species in anagenesis. *American Naturalist*, **130**, 465–473.

Gaggiotti OE, Bekkevold D, Jorgensen HBH *et al.* (2009) Disentangling the effects of evolutionary, demographic, and environmental factors influencing genetic structure of natural populations: Atlantic herring as a case study. *Evolution*, **63**, 2939–2951.

Goncalves H, Martinez-Solano I, Pereira RJ, Carvalho B, Garcia-Paris M, Ferrand N (2009) High levels of population subdivision in a morphologically conserved Mediterranean toad (*Alytes cisternasii*) result from recent, multiple refugia: evidence from mtDNA, microsatellites and nuclear genealogies. *Molecular Ecology*, **18**, 5143–5160.

Griffiths AM, Sims DW, Cotterell SP *et al.* (2010) Molecular markers reveal spatially segregated cryptic species in a critically endangered fish, the common skate (*Dipturus batis*). *Proceedings of the Royal Society B – Biological Sciences*, **277**, 1497–1503.

Hey J (2006) Recent advances in assessing gene flow between diverging populations and species. *Current Opinion in Genetics & Development*, **16**, 592–596.

Hey J, Nielsen R (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*, **167**, 747–760.

Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, **9**, 1322–1332.

Jansson R, Dynesius M (2002) The fate of clades in a world of recurrent climatic change: Milankovitch oscillations and evolution. *Annual Review of Ecology and Systematics*, **33**, 741–777.

Judkins DC, Wright R (1974) New records of mysids Boreomysis-Nobilis G O Sars and Mysis-Litoralis (Banner) in Saguenay Fjord (St. Lawrence Estuary). *Canadian Journal of Zoology*, **52**, 1087–1090.

Longhurst A (1998) *Ecological Geography of the Sea*. Academic Press, San Diego, CA.

McKinnon GE, Smith JJ, Potts BM (2010) Recurrent nuclear DNA introgression accompanies chloroplast DNA exchange

between two eucalypt species. *Molecular Ecology*, **19**, 1367–1380.

Mila B, Carranza S, Guillaume O, Clobert J (2010) Marked genetic structuring and extreme dispersal limitation in the Pyrenean brook newt *Calotriton asper* (Amphibia: Salamandridae) revealed by genome-wide AFLP but not mtDNA. *Molecular Ecology*, **19**, 108–120.

Murtskhvaladze M, Gavashelishvili A, Tarkhnishvili D (2010) Geographic and genetic boundaries of brown bear (*Ursus arctos*) population in the Caucasus. *Molecular Ecology*, **19**, 1829–1841.

Nielsen EE, Hemmer-Hansen J, Larsen PF, Bekkevold D (2009a) Population genomics of marine fishes: identifying adaptive variation in space and time. *Molecular Ecology*, **18**, 3128–3150.

Nielsen EE, Hemmer-Hansen J, Poulsen NA *et al.* (2009b) Genomic signatures of local directional selection in a high gene flow marine organism; the Atlantic cod (*Gadus morhua*). *BMC Evolutionary Biology*, **9**, 276.

Nosil P, Funk DJ, Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, **18**, 375–402.

O'Hanlon PC, Peakall R (2000) A simple method for the detection of size homoplasy among amplified fragment length polymorphism fragments. *Molecular Ecology*, **9**, 815–816.

Peakall R, Smouse PE (2006) GENALEX6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes*, **6**, 288–295.

Peccoud J, Ollivier A, Plantegenets M, Simon J (2009) A continuum of genetic divergence from sympatric host races to species in the pea aphid complex. *Proceedings of the National Academy of Sciences*, **106**, 7495–7500.

Perez-Figueroa A, Garcia-Pereira MJ, Saura M, Rolan-Alvarez E, Caballero A (2010) Comparing three different methods to detect selective loci using dominant markers. *Journal of Evolutionary Biology*, **23**, 2267–2276.

Praebel K, Westgaard JI, Fevolden SE, Christiansen JS (2008) Circumpolar genetic population genetic structure of capelin (*Mallotus villosus*). *Marine Ecology Progress Series*, **360**, 189–199.

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

Spinks PQ, Thomson RC, Shaffer HB (2010) Nuclear gene phylogeography reveals the historical legacy of an ancient inland sea on lineages of the western pond turtle, Emys marmorata in California. *Molecular Ecology*, **19**, 542–556.

Vekemans X (2002) *AFLP-SURV*. Laboratoire de Genetique et Ecologie Vegetale, Universite Libre de Bruxelles, Belgium.

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

Wakeley J, Hey J (1997) Estimating ancestral population parameters. *Genetics*, **145**, 847–855.

White TA, Stamford J, Hoelzel AR (2010) Local selection and population structure in a deep-sea fish, the roundnose grenadier (*Coryphaenoides rupestris*). *Molecular Ecology*, **19**, 216–226.

Williams LM, Oleksiak MF (2008) Signatures of selection in natural populations adapted to chronic pollution. *BMC Evolutionary Biology*, **8**, 282.

Yang DS, Kenagy GJ (2009) Nuclear and mitochondrial DNA reveal contrasting evolutionary processes in populations of deer mice (*Peromyscus maniculatus*). *Molecular Ecology*, **18**, 5115–5125.

Zhivotovsky LA (1999) Estimating population structure in diploids with multilocus dominant DNA markers. *Molecular Ecology*, **8**, 907–913.

G.J.C. is fascinated with the evolution of different phenotypes, sexual signals and behaviors within and among populations. G.J.C.'s research focuses on the evolutionary history of gene flow among populations, the evolution of phenotypic and/or behavioral differences among and within populations, and the consequences of those differences for divergence and speciation.

Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 Assignment rates and error rates of AFLPOP hybrid allocations based on equal sample sizes (one sample site) in each clade.

Table S2 Pairwise F_{ST} comparisons among all sites of the NWA using only neutral loci.

Table S3 Results of analysis of associations between genetic structure and environmental variables.

Fig. S1 Results of Structure analysis without site as LOCPRIOR.

Fig. S2 Results of outlier analysis on sites within the NWA using the FDIST (Beaumont & Nichols 1996) approach implemented in Arlequin 3.5 (Excoffier & Lischer 2010).

Fig. S3 Estimating the number of genetic clusters (*K*) with the methodologies of (a) Evanno *et al.* (2005), and (b) Pritchard *et al.* (2000).

Fig. S4 PC graphs of Northwest Atlantic sample sites, (a) with all data, (b) without four outliers.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.