Implementation of mixture analysis on quantitative traits in studies of neutral versus selective divergence

Pertoldi, Cino; Jørgensen, Hanne Birgitte Hede; Randi, Ettore; Jensen, Lasse Fast; Kjærsgaard, Anders; Loeschcke, Volker; Faurby, Søren

Published in:
Evolutionary Ecology Research

Publication date:
2012

Document Version
Early version, also known as pre-print

Evolutionary Ecology Research, 2012, **14**: 881–895

Implementation of mixture analysis on quantitative traits in studies of neutral versus selective divergence

Cino Pertoldi[1,2,3], Hanne Birgitte Hede Jørgensen[4], Ettore Randi[2,5], Lasse Fast Jensen[6], Anders Kjærsgaard[1], Volker Loeschcke[1] and Søren Faurby[1,7]

[1]*Department of Bioscience, Aarhus University, Aarhus, Denmark,* [2]*Department 18/Section of Environmental Engineering, Aalborg University, Aalborg, Denmark,* [3]*Aalborg Zoo, Aalborg, Denmark,* [4]*Department of Molecular Biology and Genetics, Aarhus University, Tjele, Denmark,* [5]*Laboratorio di Genetica, Istituto Superiore per la Protezione e la Ricerca Ambientale, Ozzano Emilia (BO), Italy,* [6]*Fisheries and Maritime Museum, Esbjerg, Denmark and* [7]*Department of Ecology and Evolutionary Biology, University of California, Los Angeles, USA*

ABSTRACT

Background: The spatial genetic structuring of natural populations is mostly studied using neutral markers. Recently, morphometric methods have also been used to to study genetic divergence through adaptive processes. These methods provide better insights into the conservation needs of focal populations. However, all morphometric methods assume that samples obtained in different localities represent distinct populations when, in fact, they may constitute a mixture of several populations due to cryptic population structure and/or environmental variability. This may lead to biased estimates of the adaptive divergence between populations. Mixture analysis makes no *a priori* assumption of the affiliation of samples. It can therefore be used to assign samples and detect population structure, allowing estimation of morphometric divergence.

Methods: We perform mixture analyses on simulated data to estimate potential bias in adaptive population divergence measures due to *a priori* assumptions about the population structure. We present three examples illustrating the possible uses of mixture analyses for identification of distinct compartments (groups of individuals that are morphologically similar) between and within populations.

Key assumptions: We assume that the presence of distinct compartments between populations can be attributed to different environmental conditions, the presence of barriers reducing gene flow, and phylogenetic signals and plasticity of the traits analysed.

Conclusions: Certain cases of (cryptic) population structure may lead to substantial bias in the estimation of population morphometric divergence. This can have major implications for conservation guidelines and for the detection of evolutionarily distinct populations.

Keywords: evolutionarily significant unit, F_{ST}-Q_{ST} and F_{ST}-P_{ST} comparisons, genetic structure, local adaptation, morphometrics, natural selection.

Correspondence: C. Pertoldi, Department of Bioscience, Aarhus University, Ny Munkegade 114, 8200 Aarhus, Denmark. E-mail: biocp@nf.au.dk

INTRODUCTION

Landscape genetics combine population genetics, spatial statistical analyses, and landscape ecology. It seeks to unravel the interactions between features of the landscape and microevolutionary processes and identify spatial barriers to gene flow (Manel *et al.*, 2003; Coop *et al.*, 2010). Thus the principles of landscape genetics are central to conservation genetics. It plays an increasingly important role in the management and conservation of species due to the need to evaluate the effects of habitat degradation and fragmentation (Manel *et al.*, 2003).

The spatial structuring of natural populations across landscapes is the product of demographic factors, including gene flow as well as random genetic drift through finite population sizes and environmental factors. Neutral molecular genetic markers such as microsatellite DNA, single nucleotide polymorphisms (SNPs), and mitochondrial DNA have been used extensively within the field of conservation genetics. They have been used to elucidate and quantify the spatio-temporal distribution of genetic variance, estimate demographic parameters as well as designate biological entities of special concern, i.e. evolutionarily significant units and conservation units (Waples, 1991; Fraser and Bernatchez, 2001). While designation of evolutionarily significant units or conservation units aims to preserve significant biological legacy and allow the potential for future adaptive evolution, a number of different approaches have been proposed. Some of these focus on divergence of allele frequencies at nuclear loci and reciprocal monophyly of mitochondrial DNA (Moritz, 1994). Based entirely on neutral genetic markers, these approaches may only be appropriate for providing insight into adaptive variation when a large fraction of the putatively neutral loci are tightly linked to quantitative trait loci of phenotypic traits under selection or when the population in question is small. In the latter case most variation, including quantitative variance, is expected to behave neutrally due to extensive random genetic drift (Pertoldi and Bach, 2007). A broader definition of evolutionarily significant units and conservation units including non-neutral markers or quantitative traits would therefore be more appropriate in other circumstances (Crandall *et al.*, 2000; Fraser and Bernatchez, 2001; Fabiani *et al.*, 2003). This will provide further information about adaptive evolutionary processes.

Knowledge about local adaptation and adaptive potential of natural populations is becoming increasingly relevant due to anthropogenic changes to the environment, including climate change. Divergent natural selection due to spatially varying environments is expected to promote adaptive evolutionary responses (Kawecki and Ebert, 2004). However, when populations are small or gene flow is extensive, populations are expected to be neutrally differentiated or genetically homogeneous, respectively. Hence, the evolutionary outcome is dictated by the relative strength of natural selection, migration, and gene flow (Endler, 1986). Selective forces influence populations in various parts of a species distribution differently (Andersson, 1994) and, in a given population, the degree of adaptation is the residual effect of the dynamic interaction between the selective pressure and gene flow. While natural selection is a potent force driving population differentiation and determining phenotypic diversity in natural populations, its importance relative to random genetic drift remains unclear (Edelaar *et al.*, 2011). The extent to which adaptive responses should be held responsible for the patterns of biological diversity is far from settled.

Whereas some insights can be obtained by consideration of rates of gene flow and migration based on putatively neutral molecular markers, direct demonstration of local adaptation involves either comparison of fitness among populations in local

and foreign environments, analysis of genes subject to selection or evaluation of the between-population component of additive genetic variance of quantitative traits (Endler, 1986; Kawecki and Ebert, 2004; Jensen *et al.*, 2008). A popular approach to unravel the relative effects of neutral evolutionary processes (i.e. random genetic drift and gene flow) and adaptive processes is to contrast population divergence at putatively neutral molecular markers (F_{ST}) with divergence in quantitative variation (Q_{ST}) or phenotypic divergence (P_{ST}) of morphological, behavioural or life-history traits assumed to be under the additional influence of natural selection (Lande, 1992; Lynch, 1996; Merilä and Crnokrak, 2001; McKay and Latta, 2002; Jensen *et al.*, 2008; Brommer, 2011). Thus, estimates of neutral genetic variation provide a null hypothesis or neutral expectation to the alternative hypothesis of adaptive divergence (Spitze, 1993; Schluter, 2001; Jensen *et al.*, 2008). When considering the relation between Q_{ST} and F_{ST}, or P_{ST} and F_{ST}, three scenarios are possible. First, a higher divergence in quantitative traits compared with neutral molecular markers ($Q_{ST} > F_{ST}$ or $P_{ST} > F_{ST}$) indicates directional selection among populations. Second, the opposite scenario (i.e. $Q_{ST} < F_{ST}$ or $P_{ST} < F_{ST}$) suggests that the same genotypes are favoured in different populations due to stabilizing selection. Third, if the two measures do not differ significantly, the possibilities of genetic drift versus selection cannot be disentangled. Despite the fact that these kinds of comparative approaches are quite promising, we should bear in mind that for many species, especially those that are endangered or vulnerable, estimation of Q_{ST} is not possible. The estimation of Q_{ST} (i.e. estimation of the additive genetic component of a phenotypic trait) requires complex experimental designs in which the environmental conditions can be manipulated. Also, the relationships between individuals must be known. In contrast, P_{ST} has often been used as a coarse surrogate of Q_{ST}. P_{ST} does not require any assumptions about controlled environmental conditions and known relationships between individuals. However, how well the P_{ST} value approximates Q_{ST} is determined by the relative importance that the additive genetic variance has in determining the between- and within-population phenotypic variation. Clearly, environmental factors, genotype × environment, and non-additive genetic variances bias such an approximation (Brommer, 2011).

Although contrasting selectively neutral and adaptive divergence is a highly useful approach when investigating the biological significance of trait variance in a conservation context, there are some caveats that need to be kept in mind during interpretation. Being under the influence of both environmental and genetic effects as well as interaction and covariance terms, estimating P_{ST} or Q_{ST} for quantitative traits potentially introduces bias, if environmental heterogeneities are not eliminated (Brommer, 2011). This is a concern when estimating P_{ST} or Q_{ST} from quantitative traits measured in the field, but environmentally induced bias in studies under controlled environments cannot be ruled out due to potential ontogenetic effects on individuals acclimated, maternal effects, and uncontrolled micro-environmental heterogeneities within the controlled environment, even if some experimental designs can partly control for some of these biases (Pujol *et al.*, 2008). As previously mentioned, the approximation that P_{ST} is equal to Q_{ST} is debatable. However, making a few assumptions in Equation (2) of Brommer (2011) for the estimation of P_{ST}, it is possible to validate such an approximation. That is, if we assume that we know the value of the scalar c (the proportion of the total phenotypic variance assumed to be caused by additive genetic effects) and of the heritability h^2 of the trait studied. To simplify the methodological part of this study, we assumed that P_{ST} is equal to Q_{ST}, and therefore we will only refer to Q_{ST} in the remainder of the paper. However, this assumption will not affect our conclusions, which are valid both for estimates of both P_{ST} and Q_{ST}.

Another issue is that sampled individuals from the wild are most often categorized into populations by making *a priori* assumptions about population of origin, typically supported by the geographical co-location of the populations. There is, however, a risk that mis-assigned migrants, based on sampling site, can lead to underestimated population differentiation and biased conclusions. Clearly, such a problem can be resolved if molecular markers are available for the species studied, as a pool of markers allows for the assignment of an individual to the population of origin. Molecular markers are, however, not always available. Cryptic population structure can also introduce significant bias. Variation in quantitative traits among different cohorts due to fluctuating environmental conditions or selective pressures leads to underestimation of P_{ST} or Q_{ST}. The same applies to cryptic spatial heterogeneities within the geographical range of *a priori* defined populations.

Cryptic population structure or erroneous *a priori* assumptions about populations can potentially be unravelled by clustering individuals. Several procedures to determine genetic population structure based on molecular genetic data without the *a priori* definition of existing populations are available and implemented in landscape genetic software [e.g. STRUCTURE (Pritchard *et al.*, 2000), BAPS (Corander *et al.*, 2008), and TESS (Durand *et al.*, 2009)]. Although the literature is rich with studies using these types of software (e.g. Mucci *et al.*, 2010), very few have attempted to elucidate population structure using mixture analysis of quantitative data (e.g. morphometric, life-history trait or gene expression studies).

The fitting of normal or t-component mixture models to multivariate data, using maximum likelihood via the EM algorithm, is widely adopted (McLachan and Krishnan, 1997; McLachan and Peel, 1998). A major advantage of mixture analysis is that, unlike many other approaches, it performs an unbiased analysis of the data without any *a priori* expectations. For this reason, Mariott (1974) dubbed it 'the only clustering process that is entirely mathematically justifiable'. The method assumes that the data are composed of a mixture of several compartments and splits the data into these clusters. No geographical information is used with this method and the grouping of a significant amount of individuals from the same localities in the same cluster therefore provides strong evidence of a geographic differentiation. Even though mixture analyses on biological data have a long history (Pearson, 1894), they have only been used in a few biological studies (Airoldi *et al.*, 1995; Pertoldi *et al.*, 2006, 2009, 2012; Faurby *et al.*, 2011). Pertoldi *et al.* (2006) conducted a morphometric study on skulls of the Iberian lynx *Lynx pardinus*, using univariate, multivariate, and mixture analysis approaches. All three techniques provided evidence for morphometric differentiation, both in skull size and shape, among three populations of geographically separated populations. Pertoldi *et al.* (2009) conducted a morphometric study followed by mixture analyses on skull traits and teeth traits of polar bear *Ursus maritimus* skulls sampled in East Greenland from 1892 to 2002. The mixture analyses, followed by multivariate analyses, provided evidence for morphometric differences in both the size and the shape of individual skulls collected. The fact that environmental and genetic changes produce different combinations of patterns of morphometric changes allowed the authors to individuate the causes of the morphometric modifications. Faurby *et al.* (2011) analysed shape variation in different species of Horseshoe crabs (*Limulus polyphemus* and *Carcinoscorpius rotundicauda*), and by comparing the degree of geographic variation between sexes and species found strong indications for the importance of both sexual and natural selection.

The concomitant use of genetic markers for the detection of genetic differentiation together with analyses of variance at quantitative traits can provide a more detailed answer to the questions that arise when deciding a conservation strategy (e.g. reintroduction,

translocation or repopulation). In this study, using computer simulations, we investigate the potential use of mixture analyses on quantitative data in a conservation and landscape genetic context. Specifically, we elucidate the power of mixture analyses to pick up signals of cryptic population structure and investigate the introduced bias in estimates of quantitative divergence (Q_{ST}), when cryptic structure is present. We discuss potential applications on a suite of quantitative data (univariate or multivariate), including morphometric (metric and meristic), demographic, gene-expression, physiological, and environmental data. We also provide suggestions as to how mixture analyses of quantitative data can be implemented in the landscape genetics software that is routinely used to determine population genetic structure and geographical distribution of population clusters simultaneously by using information on genotypes and geographic locations.

METHODS

Computational issues

To determine the bias produced when considering a sample that is a mixture of different distributions and to test the capacity to detect such an admixture, we ran a number of simulations in R v.2.11.1 using functions from the package mixtools and custom codes (Benaglia *et al.*, 2009; R Development Core Team, 2010). All simulations were run assuming two clusters and mixture analyses were set to search for mixtures of normal distributions.

In these simulations we estimated biases on Q_{ST} estimates when, in reality, an assumed homogeneous population consists of two populations whose trait measurements show separate normal distributions. First, we estimated the average bias on Q_{ST} when the two trait distributions have different mean values. Second, we estimated the average bias on Q_{ST} when the two trait distributions have identical means but different variances around the mean. Third, we estimated the average bias on Q_{ST} estimates when individuals are assigned to populations based on measurements of one or more traits (and with different phenotypic correlations between the traits).

In all the simulations, we compared two populations with sample size 100 and estimated Q_{ST} as (Between-group SS/(Between-group SS + 2 Within-group SS)), where SS is sum of squares. For all analyses, we compared our mixed distribution with a basic distribution P_{Basic} that had a normal distribution, a mean of 10, and standard deviation of 1/3.

Population mixture composed of distributions with different means

To test for the effect that a mixture of two normal distributions with different means can have on the Q_{ST} estimates, we estimated Q_{ST} values between the basic population, P_{Basic}, and a mixed population, P_{Mix}, with an overall mean varying between 8 and 12 and with standard deviation of 1/3. The mixed distributions consisted of a mixture of trait measurements that were known to originate from one of two populations with different sample size ($P_{SmallKnown}$ and $P_{BigKnown}$). For each simulation, we generated the distribution of trait measurements for P_{Mix} by sampling varying proportions of each of the $P_{SmallKnown}$ and $P_{BigKnown}$ distributions. Specifically, we sampled between 5 and 50 individuals from $P_{SmallKnown}$ and the remaining individuals from $P_{BigKnown}$ to give a total population size of 100 individuals in P_{Mix}. The mean trait value of $P_{SmallKnown}$ was set to $((P_{SmallKnown} + P_{BigKnown})/100) + 1$, while the mean trait value of $P_{BigKnown}$ was adjusted accordingly to keep the overall mean constant.

For each of the nine different means produced by mixing $P_{SmallKnown}$ and $P_{BigKnown}$ with different proportions of the mixed distributions (8, 8.5 ... 12), tests were run for ten different sizes of the small population (5, 10 ... 50), generating a total of 90 different distributions. For each P_{Mix}, we ran a mixture analysis and assigned individuals to $P_{SmallAssign}$ or $P_{BigAssign}$. $P_{SmallAssign}$ and $P_{BigAssign}$ were defined so that the ratio between the means of the two mixtures was the same as the ratio between the means for $P_{SmallKnown}$ and $P_{BigKnown}$. After this step, we sampled with repeat from $P_{SmallKnown}$, $P_{BigKnown}$, $P_{SmallAssign}$, and $P_{BigAssign}$ to generate four populations with sample size 100 each, since Q_{ST} is influenced by the ratio of sample sizes between populations. Finally, we calculated Q_{ST} values between P_{Basic} and the four populations $P_{SmallKnown}$, $P_{BigKnown}$, $P_{SmallAssign}$, and $P_{BigAssign}$, as well as Q_{ST} values between P_{Basic} and P_{Mix}. To remove any potential sampling error effect, we sampled ten times from the results of each mixture analysis and only considered the mean of the Q_{ST} values calculated from these ten replicates. In some cases, the mixture analysis assigned all individuals to a single component. These cases were ignored for the purpose of calculating Q_{ST} for P_{Mix}, $P_{SmallKnown}$, and $P_{BigKnown}$, while the Q_{ST} for $P_{SmallAssign}$ was defined as 1 and the Q_{ST} for $P_{BigAssign}$ was defined as the median Q_{ST} for P_{Mix} for the set of 100 simulations in question.

Since we were only interested in average effects, all analyses focused on median values for each distribution of overall means of the mixed population. The median was chosen over the mean, since Q_{ST} values will have asymmetrical errors as they only range between 0 and 1, making the median values more informative.

Population mixture composed of distributions with different variances

To test for the effect that a mixture of trait measurements from two normal distributions with identical means but different variances can have on Q_{ST}, we ran analyses with basic set-ups as above but with the following adjustments: Both $P_{SmallKnown}$ and $P_{BigKnown}$ had the same mean (separate analyses were run for means equal to 10.5 and 11) but whereas the standard deviation of $P_{BigKnown}$ was kept constant (at 1/3), the standard deviation of $P_{SmallKnown}$ varied between 1/12 and 4/3. In these cases, the measurement distributions for $P_{SmallAssign}$ and $P_{BigAssign}$ were defined to ensure that the ratio between the standard deviations of the two mixtures was the same as the ratio between the standard deviations for $P_{SmallKnown}$ and $P_{BigKnown}$.

Q_{ST} bias estimation

To assess the bias in Q_{ST} estimates based on single trait measurements versus measurements on more than one trait, we assigned individuals to two populations based on measurements on one, two, and three traits. Bias is presented as the differences between estimated Q_{ST} values (based on assignments from mixture analyses) and the true (known) Q_{ST} estimates. Since the bias depends on the size of the populations, we focused on the difference between Q_{ST} for $P_{SmallKnown}$ and $P_{SmallAssign}$ and analysed them for univariate, bivariate, and trivariate normally distributed data. For the bivariate and trivariate data sets, we analysed multivariate distributions with low ($\rho = 0.2$), medium ($\rho = 0.4$), and high ($\rho = 0.8$) correlations between trait measurements. For analyses based on bivariate and trivariate measurement distributions, only one of the univariate distributions was used to calculate Q_{ST}, while the other distributions were used for assignment in the mixture analyses.

For the bivariate and trivariate data sets, we assigned the data by performing univariate mixture analyses for each variable and calculating the probability that individuals belonged

to $P_{SmallAssign}$ based on measurements for one, two, and three traits. These probabilities are referred to as $Prob_{Small\_1}$, $Prob_{Small\_2}$, and $Prob_{Small\_3}$. Assignments were based on mean values or standard deviations and calculated as

$$\frac{Prob_{Small\_1} \times Prob_{Small\_2}}{Prob_{Small\_1} \times Prob_{Small\_2} + (1 - Prob_{Small\_1}) \times (1 - Prob_{Small\_2})}$$

(if the data were bivariate), or

$$\frac{Prob_{Small\_1} \times Prob_{Small\_2} \times Prob_{Small\_3}}{Prob_{Small\_1} \times Prob_{Small\_2} \times Prob_{Small\_3} + (1 - Prob_{Small\_1}) \times (1 - Prob_{Small\_2}) \times (1 - Prob_{Small\_3})}$$

(if the data were trivariate), and individuals were assigned to $P_{SmallAssign}$ if this probability was above 0.5. Following this assignment, we estimated the Q_{ST} values for $P_{SmallKnown}$ compared with P_{Basic} and for $P_{SmallAssign}$ compared with P_{Basic}.

RESULTS

Population mixture composed of distributions with different means

For mixtures composed of different means, no systematic bias in the estimation of Q_{ST} was identified, as the median Q_{ST} values for $P_{SmallAssign}$ and $P_{BigAssign}$ were nearly identical to the median Q_{ST} values for $P_{SmallKnown}$ and $P_{BigKnown}$ for all analysed means of P_{Mix} (Figs. 1a–i). Furthermore, while the difference in Q_{ST} between the mixed and largest population always increased with increasing size of P_{Small}, quite marked differences between Q_{ST} for P_{Big} and P_{Mixed} vs. P_{Basic} were observed when the size of the smallest compartment was 10–15% of the entire mixed population (Figs. 1a–i). An additional point is evident from the subplot of the mean (μ) equal to 12 (Fig. 1i), which shows that the Q_{ST} of a mixed population can be less than the Q_{ST} of each of the subgroups due to the increased variance in the mixed population.

Population mixture composed of distributions with different variances

The consequences of a mixture composed of two distributions with different variances are shown in Fig. 2. In these cases, there were systematic biases for analyses with relatively moderate (two-fold) differences in standard deviations between the components (Figs. 2b–c, f–g). The differences between the Q_{ST} values calculated for the assigned vs. the known components were greater than the differences between any of the true components and the mixed population.

These systematic differences disappeared for the larger (four-fold) differences in standard deviations (Figs. 2a, d, e, h) between the components, showing that mixture analyses are fully capable of separating populations with identical mean as long as the difference in their standard deviations are large enough. The existence of a small component with less variation appeared unimportant, since Q_{ST} for P_{Big} was close to Q_{ST} for P_{Mix} (Figs. 2a, e). The existence of a small, more varied component led to much larger deviations between Q_{ST} for P_{Big} and Q_{ST} for P_{Basic} (Figs. 2d, h). These analyses essentially showed that the Q_{ST} for P_{Mix} is mainly driven by the most variable component if the means of the components are identical.

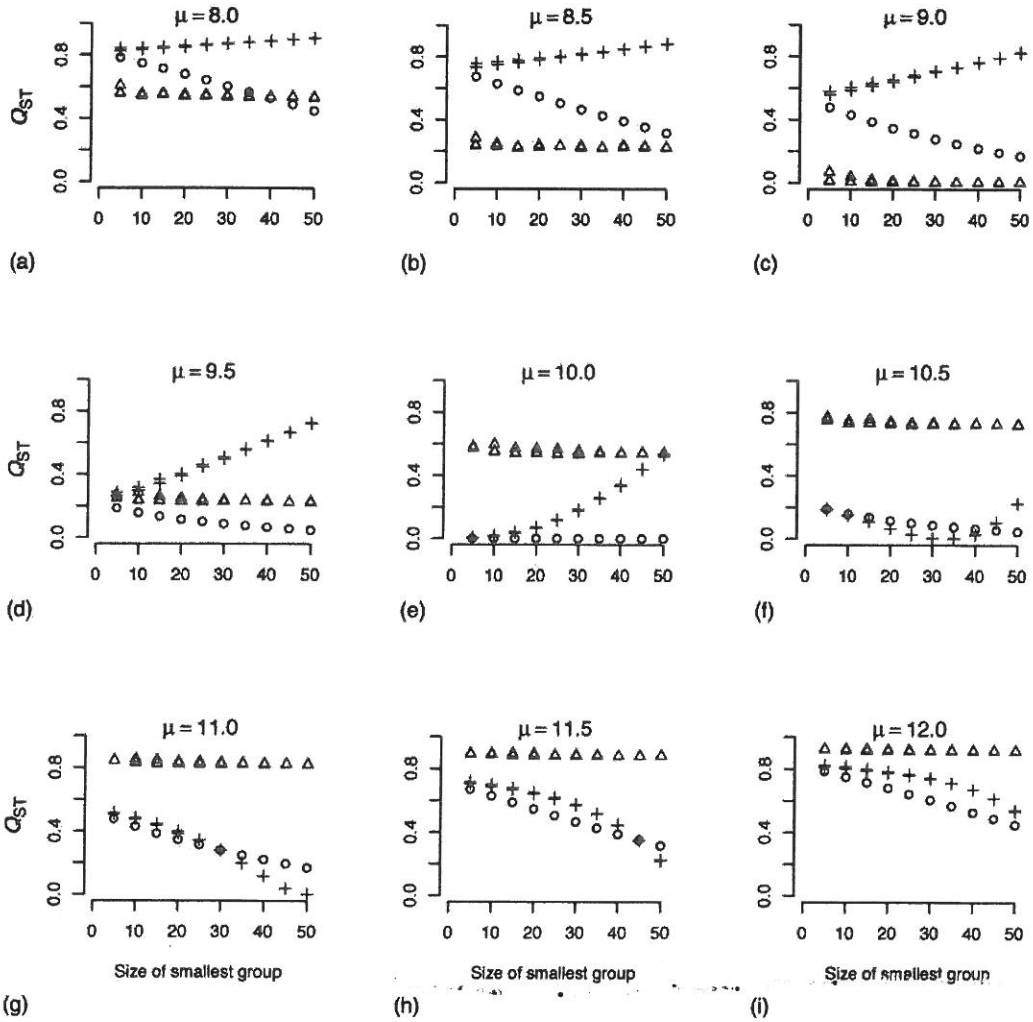

Fig. 1. Analysis of Q_{ST} calculations with different means for P_{Mix}. Nine different subplots show results of mean P_{Mix} between $\mu = 8.0$ and 12.0. Q_{ST} values between P_{Basic} and P_{Mix} are shown by circles; between P_{Basic} and P_{Big} by crosses; and between P_{Basic} and P_{Small} by triangles. Values calculated with perfect assignment are shown in black, while values calculated with assignment from the mixture analyses are shown in grey.

Q_{ST} bias estimation

Whereas the previous analyses focused on the overall pattern produced by mixtures composed of different means or different variances, a separate issue arises when investigating the amount of bias caused by non-perfect assignment to individual components (Figs. 3a–d). It is evident that the bias became substantial when P_{Small} was small. Although the size of this error consistently decreased as the proportion of P_{Small} individuals in P_{Mix} increased, the errors were minor as long as at least 15% of the individuals in P_{Mix} belonged to P_{Small} if the means of P_{Small} and P_{Big} were different or if the variance of P_{Small} was less

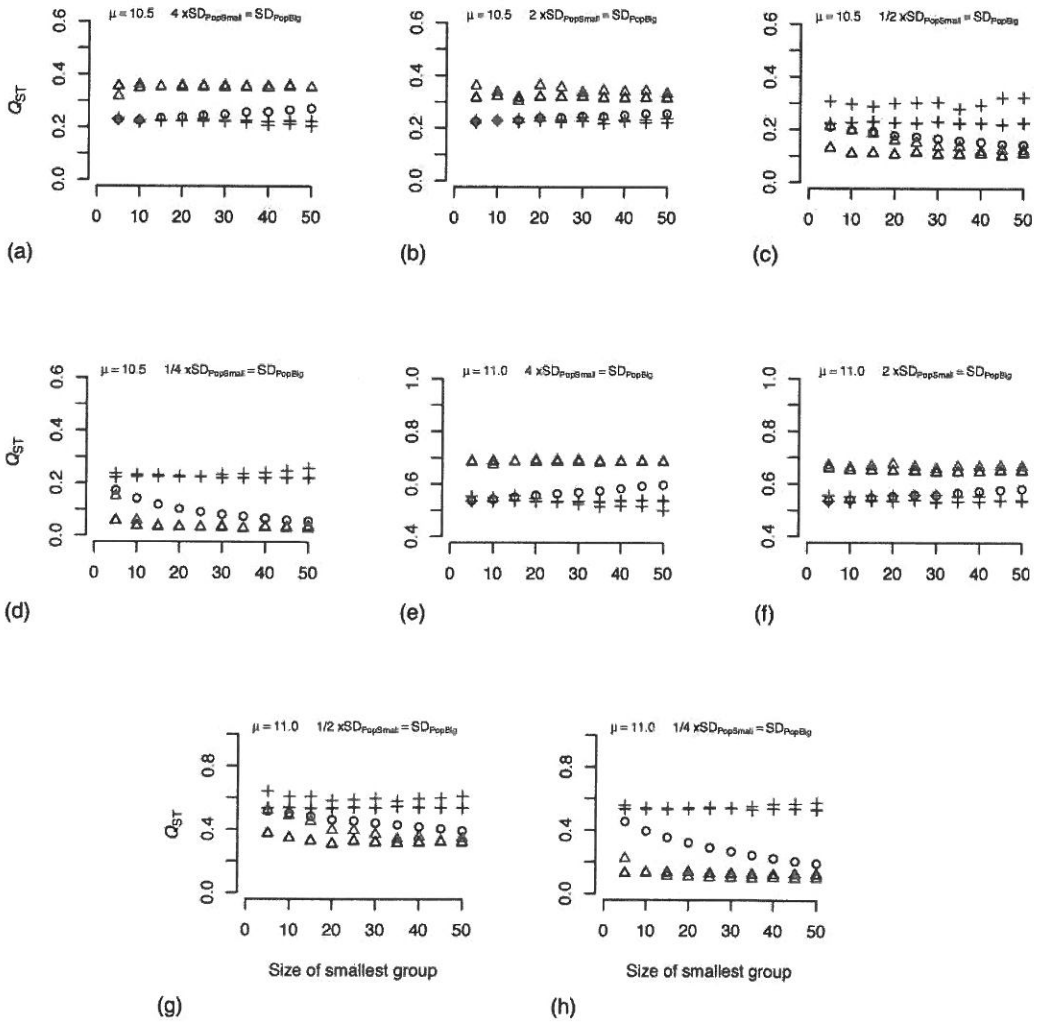

Fig. 2. Analysis of Q_{ST} calculations with different standard deviations for each component of P_{Mix}. Eight different subplots show results of differences in standard deviation of P_{Small} vs. P_{Big} calculated with mean P_{Big} of either $\mu = 10.5$ or 11.0. Q_{ST} values between P_{Basic} and P_{Big} are shown by crosses and those between P_{Basic} and P_{Small} by triangles. Values calculated with perfect assignment are shown in black, while values calculated with assignment from the mixture analyses are shown in grey.

than that of P_{Big} (Figs. 3a, b, d). For the final scenario analysed, a P_{Small} with a higher variance than P_{Big}, around 30% of the individuals had to belong to P_{Small} to obtain a fairly reliable measurement of Q_{ST} between P_{Small} and P_{Basic} (Fig. 3c).

It is also clear from these analyses that the errors for the univariate data set are substantially larger than the errors for the bivariate or trivariate data set. Among these multivariate data sets, the errors are smaller for data with a low correlation between the parameters ($\rho = 0.2$) than for moderate ($\rho = 0.4$) or high correlations ($\rho = 0.8$) (Fig. 3). The difference between two dimensions (full lines) and three dimensions (points) were much less

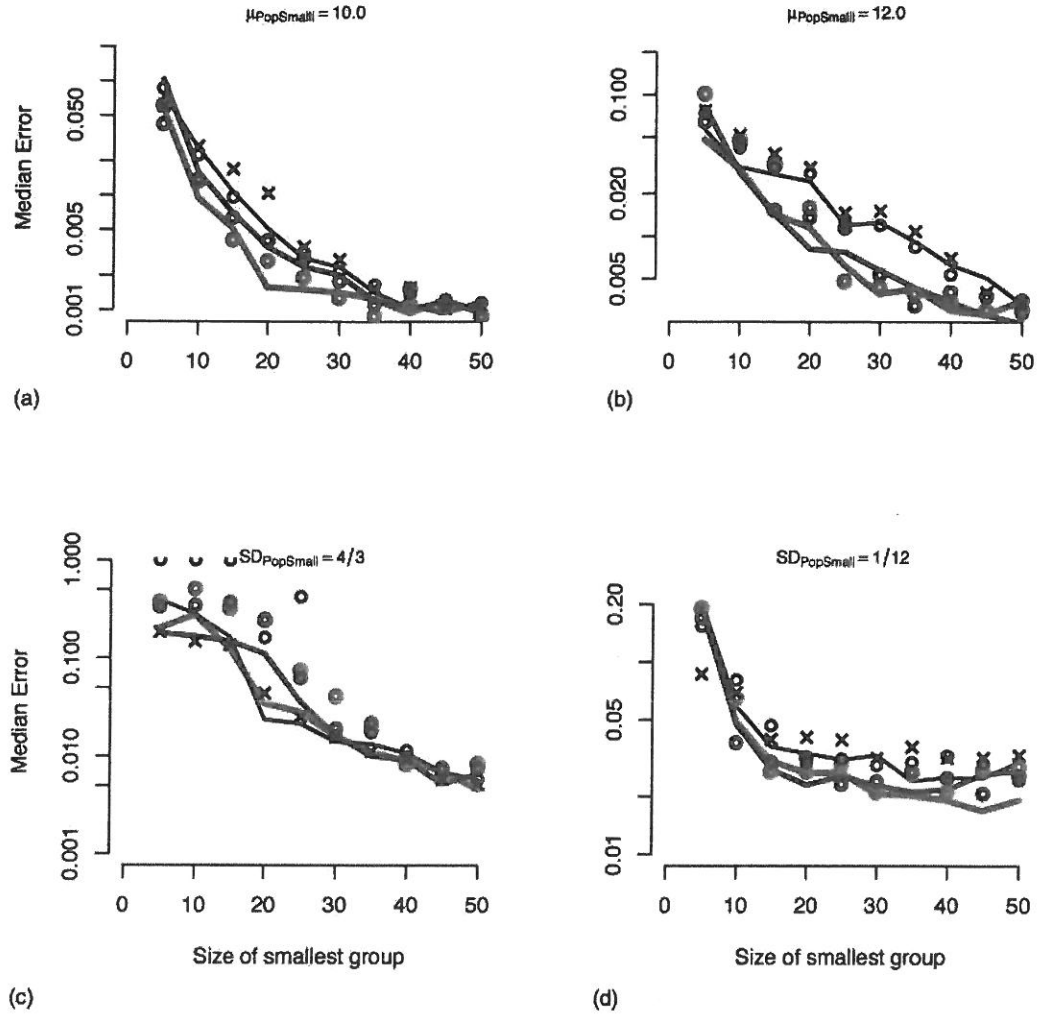

Fig. 3. Measurement errors in Q_{ST} by imperfect assignment of mixture analyses. The median difference between Q_{ST} for $P_{SmallKnown}$ and $P_{SmallAssign}$ in each simulation calculated under four different scenarios. The first two scenarios are represented in the third and ninth subplots of Fig. 1 (a and i), while the last two scenarios are represented in the fifth and eight subplots of Fig. 2 (e and h). The crosses represent data for the univariate data set; the grey lines represent data for the bivariate data set; and the circles represent data for the trivariate data set. Results for low correlated multidimensional data ($\rho = 0.2$) are shown by the thickest lightest grey line, for moderate correlated multidimensional data ($\rho = 0.4$) by the line of intermediate thickness, and for highly correlated data ($\rho = 0.8$) by the thin dark grey line.

evident, suggesting that the correlation between the parameters was much less important that the number of correlations.

Our results suggest that Q_{ST} estimates are highly influenced by the presence of non-homogeneous means but that such mixtures generally can be reliably unmasked with mixture analyses making the problem fairly easy to handle. Non-homogeneous variance

is harder to handle and moderate differences in variance between sub-compartments are potentially better ignored, as the bias they introduce may be smaller than the errors caused by non-perfect assignment to clusters.

If the variance of P_{Small} is substantially lower than the variance of P_{Big}, mixture analyses can unmask the situation but the Q_{ST} between P_{Big} and P_{Basic} is very close to the Q_{ST} between P_{Mix} and P_{Basic} in such situations and mixture analyses may not be vital. The existence of a P_{Small} with a substantially higher variance than P_{Big} is the most problematic situation. In such cases, the Q_{ST} between P_{Big} and P_{Basic} is very different from the Q_{ST} between P_{Mix} and P_{Basic} but the individual mixtures are harder to identify. Such situations are potentially best handled by performing mixture analyses but only analysing the data from P_{Big}.

DISCUSSION

Mixture techniques as a complementary tool in landscape genetics

We have shown that mixture techniques can be a powerful tool to get a real impression of the complex interplay between genotype and environment that shapes traits across landscapes. In this study, we have demonstrated that considering a population to be homogeneous or assigning individuals to different populations using geographic data as a criterion for the assignment can generate substantial bias when quantifying morphometric differentiation using Q_{ST}. Clearly, the other univariate or multivariate indices of morphometric distances will also be biased, although for clarity and brevity we did not investigate the bias produced by mixtures on these indices. There is no reason to expect that the same mixture technique should not apply for studies estimating population differentiation from quantitative trait data generated by transcriptomics, proteomics or metabonomics (e.g. Whitehead and Crawford, 2006). In the future, mixture analysis has the potential to provide insight into geographic variation whether caused by population history or selective forces. The cluster patterns described by the output of the mixture analysis could reveal patterns of phylogenetic signals that are illustrating history and not ecology. Several studies have reported this pattern, which appears to be common in newer splits such as studies analysing intraspecific variation or recent speciation (Macholán, 2006).

Implementation of mixture analysis could prove valuable in long-term monitoring programmes by revealing clustering into different time periods. This includes different cohorts having experienced different environmental conditions. On different geographical scales the mixture analysis could also become a complementary tool for the individuation of evolutionarily significant units and conservation units. In fact, the compartments produced by the assignment of the different individuals by the mixture analysis can be compared with the clusters produced by software traditionally used in the landscape genetics field. The potential discrepancies between the clusters produced by the software allow several interpretations. For example, the presence of two or more morphometric clusters within a group of individuals, which are indicated as one single cluster by the traditional landscape genetics software, could indicate subtle genetic substructure or environmental differences occurring in the population area of distribution. One must, however, bear in mind that spatial and temporal variation in habitat quality and population density can also affect trait size (Holbrook, 1982). The detection of spatio-temporal changes in size and shape could therefore reveal ecological patterns produced by rapid environmental

changes or even change in the genetic composition of the population due to strong demographic changes or mixtures with other populations.

Future studies of the concomitant screening of neutral and quantitative traits with the use of mixture analysis, could also add considerably to the debate on the accepted paradigm which more or less states that geographic isolation is the main determinant of population divergence (Futuyma and Mayer, 1980; Felsenstein, 1981). The effect of gene flow as a homogenizing factor that antagonizes both drift and selection has been strongly emphasized historically (Gillespie and Turelli, 1989; Stanton and Galen, 1997). However, several phenomena related to natural selection can cause the divergence of populations in the absence of geographic isolation (Ehrlich and Raven, 1969; Rice and Salt, 1990; Rice and Hostert, 1993; Schluter, 2001) and the classical paradigm needs to be re-evaluated. In fact, gene flow has often been considered responsible for preventing differentiation of populations under selection, otherwise generating local adaptation (Storfer and Sih, 1998; Lenormand, 2002). However, strong selection pressures can counterbalance its homogenizing effects (Mopper, 1996), as immigrant genes may not establish and the population under selection may remain genetically distinct in the face of migration (Nagy and Rice, 1997).

Utilization of mixture analysis on gene-expression, ecological, demographic, and physiological data

In this study, we decided to only simulate normal distributions or mixtures of normal distributions so as to simplify interpretation of the results. However, it may be possible to apply mixture analysis to non-normally distributed data. The possibility of using mixture analysis when working with, for example, non-Gaussian distributions will considerably expand its area of use. Examples include: counting/census data, which normally follow a Poisson distribution; respiration rate/data expressed as percentages/proportion data/ratio data, which normally follow a negative binomial distribution; or population dynamic data, which normally follow a log-normal distribution. Gene expression measured as mRNA levels (as microarray or RNAseq data) are also best seen as phenotypic traits (Khaitovich *et al.*, 2006), and they are assumed normally distributed by widely used statistical packages like limma (Smyth, 2004). Since the distribution of expression levels may, in fact, not meet the assumptions, the actual distributions of the expression data should be known before running mixture analyses. Selection does not act on gene expression directly, rather on the ecological, morphological, physiological or other phenotypic traits affected by changes in expression levels. Numerous cellular processes from post-transcriptional modifications of mRNA to tissue-specific responses to external stimuli make the connections between genome and phenotypes difficult to disentangle. This complexity may be partly responsible for the different approaches to studying local adaptation. Some authors focus on genetic (genomic) variation as the basis for phenotypic variation and test for associations between alleles in genetic markers and phenotypic variation (see, for example, Storz and Wheat, 2010; Elmer and Meyer, 2011). Others focus on the possibly profound effect of variation in gene expression on phenotypic differentiation among populations (see, for example, Oleksiak *et al.*, 2002; Khaitovich *et al.*, 2006).

ACKNOWLEDGEMENTS

This study was supported in part by the Danish Natural Science Research Council (grants #11-103926, #09-065999, and #95095995 to C.P.) and the Carlsberg Foundation (grant #2011-01-0059).

REFERENCES

Airoldi, J.P., Flury, B.D. and Salvioni, M. 1995. Discrimination between two species of *Microtus* using both classified and unclassified observations. *J. Theor. Biol.*, **177**: 247–262.

Andersson, M. 1994. *Sexual Selection*. Princeton, NJ: Princeton University Press.

Benaglia, T., Chauveau, D., Hunter, D.R. and Young, D. 2009. 'mixtools': an R package for analyzing finite mixture models. *J. Stat. Soft.*, **32**: 1–29.

Brommer, J.E. 2011. Whither P_{ST}? The approximation of Q_{ST} by P_{ST} in evolutionary and conservation biology. *J. Evol. Biol.*, **24**: 1160–1168.

Coop, G., Witonsky, D., Di Rienzo, A. and Pritchard, J.K. 2010. Using environmental correlations to identify loci underlying local adaptation. *Genetics*, **185**: 1411–1423

Corander, J., Sirén, J. and Arjas, E. 2008. Bayesian spatial modelling of genetic population structure. *Comput. Stat.*, **23**: 111–129.

Crandall, K.A., Bininda-Edmonds, O.R.P., Mace, G.M. and Wayne, R.K. 2000. Considering evolutionary processes in conservation biology. *Trends Ecol. Evol.*, **15**: 290–295.

Durand, E., Jay, F., Gaggiotti, O.E. and François, O. 2009. Spatial inference of admixture proportions and secondary contact zones. *Mol. Biol. Evol.*, **26**: 1963–1973.

Edelaar, P., Burraco, P. and Gomes-Mestre, I. 2011. Comparisons between Q_{ST} and F_{ST} – how wrong have we been? *Mol. Ecol.*, **20**: 4830–4839.

Ehrlich, P.R. and Raven, P.H. 1969. Differentiation of populations. *Science*, **165**: 1228–1232.

Elmer, K.R. and Meyer, A. 2011. Adaptation in the age of ecological genomics: insights from parallelism and convergence. *Trends Ecol. Evol.*, **26**: 298–306.

Endler, J.A. 1986. *Natural Selection in the Wild*. Princeton, NJ: Princeton University Press.

Fabiani, A., Hoelzel, A.R., Galimberti, F. and Muelbert, M.M.C. 2003. Long-range paternal gene flow in the southern elephant seal. *Science*, **299**: 676.

Faurby, S., Nielsen, K.S.K., Bussarawit, S., Intanai, I., van Cong, N., Pertoldi, C. *et al.* 2011. Intraspecific shape variation in horseshoe crabs: the importance of sexual and natural selection for local adaptation *J. Exp. Mar. Biol. Ecol.*, **407**: 131–138.

Felsenstein, J. 1981. Skepticism towards Santa Rosalia, or why are there so few kinds of animals? *Evolution*, **35**: 124–138.

Fraser, D.J. and Bernatchez, L. 2001. Adaptive evolutionary conservation: towards a unified concept for defining conservation units. *Mol. Ecol.*, **10**: 2741–2752.

Futuyma, D.J. and Mayer, G.C. 1980. Non-allopatric speciation in animals. *Syst. Zool.*, **29**: 254–271.

Gillespie, J.H. and Turelli, M. 1989. Genotype–environment interactions and the maintenance of polygenic variation. *Genetics*, **121**: 129–138.

Holbrook, S. 1982. Ecological inferences from mandibular morphology of *Peromyscus maniculatus*. *J. Mammal.*, **61**: 436–448.

Jensen, L.F., Hansen, M.M., Pertoldi, C., Holdensgaard, G., Mensberg, K.-L.D. and Loeschcke, V. 2008. Local adaptation in brown trout early life-history traits: implications for climate change adaptability. *Proc. R. Soc. Lond. B*, **75**: 2859–2868.

Kawecki, T.J. and Ebert, D. 2004. Conceptual issues in local adaptation. *Ecol. Lett.*, **7**: 1225–1241.

Khaitovich, P., Enard, W., Lachman, M. and Pääbo, S. 2006. Evolution of gene expression. *Nat. Rev. Genet.*, **7**: 693–702.

Lande, R. 1992. Neutral theory of quantitative genetic variance in an island model with local extinction and colonization. *Evolution*, **46**: 381–389.

Lenormand, T. 2002. Gene flow and the limits to natural selection. *Trends Ecol. Evol.*, **17**: 183–189.

Lynch, M. 1996. A quantitative-genetic perspective on conservation issues. In *Conservation Genetics: Case Histories from Nature* (J.C. Avise and J.L. Hamrick, eds.), pp. 471–501. New York: Chapman & Hall.

Macholán, M. 2006. A geometric morphometric analysis of the shape of the first upper molar in mice of the genus *Mus* (Muridae, Rodentia). *J. Zool.*, **270**: 672–681.

Manel, S., Schwartz, M.K., Luikart, G. and Taberlet, P. 2003. Landscape genetics: combining landscape ecology and population genetics. *Trends Ecol. Evol.*, **18**: 189–197.

Mariott, F.H.C. 1974. *The Interpretation of Multiple Observations*. London: Academic Press.

McKay, J.K. and Latta, R.G. 2002. Adaptive population divergence: markers, QTL and traits. *Trends Ecol. Evol.*, **17**: 285–291.

McLachan, G.J. and Krishnan, T. 1997. *The EM Algorithm and Extensions*. New York: Wiley.

McLachan, G.J. and Peel, D. 1998. Robust cluster analysis via mixtures of multivariate t-distributions. In *Advances in Pattern Recognition* (A. Amin, D. Dori, P. Pudil and H. Freeman, eds.), pp. 658–665. Sydney, NSW: Springer.

Merilä, J. and Crnokrak, P. 2001. Comparison of genetic differentiation at marker loci and quantitative traits. *J. Evol. Biol.*, **14**: 892–903.

Mopper, S. 1996. Adaptive genetic structure in phytophagous insect populations. *Trends Ecol. Evol.*, **11**: 235–238.

Moritz, C. 1994. Defining evolutionarily significant units for conservation. *Trends Ecol. Evol.*, **9**: 373–375.

Mucci, N., Arrendal, J., Ansorge, H., Bailey, M., Bodner, M., Delibes, M. *et al.* 2010. Genetic diversity and landscape genetic structure of otter (*Lutra lutra*) populations in Europe. *Conserv. Genet.*, **11**: 583–599.

Nagy, E.S. and Rice, K.J. 1997. Local adaptation in two subspecies of an annual plant: implications for migration and gene flow. *Evolution*, **51**: 1079–1089.

Oleksiak, M.F., Churchill, G.A. and Crawford, D.L. 2002. Variation in gene expression within and among natural populations. *Nat. Genet.*, **32**: 261–266.

Pearson, K. 1894. Contributions to the mathematical theory of evolution. *Phil. Trans. R. Soc. Lond. A*, **185**: 71–110.

Pertoldi, C. and Bach, L.A. 2007. Evolutionary aspects of climate-induced changes and the need for multidisciplinarity. *J. Therm. Biol.*, **32**: 118–124.

Pertoldi, C., Garcia-Perea, R., Godoy, J.A., Delibes, M. and Loeschcke, V. 2006. Morphological consequences of range fragmentation and population decline on the endangered Iberian lynx (*Lynx pardinus*). *J. Zool.*, **268**: 73–86.

Pertoldi, C., Sonne, C., Dietz, R., Schmidt, N.M. and Loeschcke, V. 2009. Craniometric characteristics of polar bear skulls from two periods with contrasting levels of industrial pollution and sea ice extent. *J. Zool.*, **279**: 321–328.

Pertoldi, C., Sonne, C., Wiig, Ø., Baagøe, H.J., Loeschcke, V. and Bechshøft, T.Ø. 2012. East Greenland and Barents Sea polar bears (*Ursus maritimus*): adaptive variation between two populations using skull morphometrics as an indicator of environmental and genetic differences. *Hereditas*, **149**: 99–107.

Pritchard, J.K., Stephens, M. and Donnelly, P.J. 2000. Inference of population structure using multilocus genotype data. *Genetics*, **155**: 945–959.

Pujol, B., Wilson, A.J., Ross, R.I.C. and Pannel, J.R. 2008. Are Q(ST)–FST comparisons for natural populations meaningful? *Mol. Ecol.*, **17**: 4782–478.

R Development Core Team. 2010. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rice, W.R. and Hostert, E.E. 1993. Laboratory experiments on speciation: what have we learned in 40 years? *Evolution*, **47**: 1637–1653.

Rice, W.R. and Salt, G. 1990. The evolution of reproductive isolation as a correlated character under sympatric conditions: experimental evidence. *Evolution*, **44**: 1140–1152.

Schluter, D. 2001. Ecology and the origin of species. *Trends Ecol. Evol.*, **16**: 372–380.

Smyth, G.K. 2004. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet., Mol.*, **3**: Article 3.

Spitze, K. 1993. Population structure in *Daphnia obtusa*: quantitative genetic and allozymic variation. *Genetics*, **135**: 367–374.

Stanton, M.L. and Galen, C. 1997. Life on the edge: adaptation versus environmentally mediated gene flow in the snow buttercup, *Ranunculus adoneus*. *Am. Nat.*, **150**: 143–178.

Storfer, A. and Sih, A. 1998. Gene flow and ineffective antipredator behavior in a stream-breeding salamander. *Evolution*, **52**: 558–565.

Storz, J.F. and Wheat, C.W. 2010. Integrating evolutionary and functional approaches to infer adaptation at specific loci. *Evolution*, **64**: 2489–2509.

Waples, R.S. 1991. Pacific salmon, *Oncorhynchus* spp. and the definition of 'species' under the Endangered Species Act. *Mar. Fish. Rev.*, **53**: 11–22.

Whitehead, A. and Crawford, D.L. 2006. Neutral and adaptive variation in gene expression. *Proc. Natl. Acad. Sci. USA*, **103**: 5425–5430.