Heredity (2009) 103, 385–393

www.nature.com/hdy

ORIGINAL ARTICLE

Estimating gametic introgression rates in a risk assessment context: a case study with Scots pine relicts

JJ Robledo-Arnuncio[1], M Navascués[2,3], SC González-Martínez[1] and L Gil[3]
[1]*Departamento de Sistemas y Recursos Forestales, CIFOR-INIA, Carretera de La Coruña km 7.5, Madrid, Spain;* [2]*Équipe Éco-évolution Mathématique, CNRS UMR 7625 Écologie and Évolution, Université Pierre et Marie Curie, École Normale Supérieure, 46 rue d'Ulm, Paris, France and* [3]*Unidad de Anatomía, Fisiología y Genética Forestales, Departamento de Silvopascicultura, ETSI de Montes, Universidad Politécnica de Madrid, Ciudad Universitaria s/n, Madrid, Spain*

The estimation of recent gene immigration is fundamental to a wide range of evolutionary and conservation studies. In a risk assessment context, gene flow estimation procedures are needed that are both accurate and readily amenable to formal evaluation of statistical uncertainty. However, genetic methods for estimating recent migration rates that are specific and have been thoroughly evaluated are scarce. Here we use an original and straightforward maximum-likelihood method to estimate recent uniparental gametic immigration from non-local plantations into an endangered population of the Iberian relict pine variety *Pinus sylvestris* var. *nevadensis* D. H. Christ. Our approach is not intended to ascertain population membership of individuals, but rather to obtain accurate immigration rate estimates with reliable confidence limits. We found very high (~40%) pollen introgression at the seed-crop level into the Scots pine relict, and substantial (10–15%) male gametic introgression among naturally regenerated recruits. Using numerical simulation, we show that our method yields uniparental gametic immigration estimates that are expected to be virtually unbiased and usually accurate under our sampling conditions. Among four tested methods to estimate the confidence intervals for immigration estimates, the profile-likelihood method was the best, as it outperformed bootstrapping procedures and yielded coverage close to nominal limits under different sample sizes and migration rates. This study presents a method by which researchers can facilitate decision making within a gene flow risk assessment context. *Heredity* (2009) **103,** 385–393; doi:10.1038/hdy.2009.78; published online 15 July 2009

Keywords: gene flow; recent migration; long distance pollen and seed dispersal; maximum likelihood; genetic introgression

Introduction

Gene flow has contrasting implications for conservation biology. On the one hand, it is thought to hamper local adaptation in heterogeneous environments by disrupting co-adapted gene complexes and changing allele frequencies in a direction opposite to divergent natural selection ('migration load' effect). On the other hand, gene flow can reduce the deleterious effects of inbreeding and increase the genotypic variance available for selection, facilitating local adaptation ('genetic rescue' effect). Although the relative importance of these effects remains unclear, the precise level of gene flow seems critical in determining the final outcome on local adaptation, for a given demographic and selective scenario (Lenormand, 2002). In practice, any objective decision regarding management of ongoing gene flow into conservation populations should be based on accurate estimates of the realized gene immigration rates, which must be appraised along with the specific demographic and selective processes affecting the population.

Estimating gene immigration rates is especially relevant when dealing with protected populations that are potentially exposed to gene flow from artificially introduced interfertile relatives. Any scientific risk assessment protocol of non-local germplasm will require reliable contemporary gene flow estimates, sometimes over broad spatial scales, which may not be easily available. Difficulties in obtaining precise estimates of ongoing gene flow at a landscape scale are illustrated by the intense scientific debate about the presence or absence of gene flow from transgenic plantations into maize landraces in Mexico: more than 10 studies conducted in recent years have disputed various statistical and sampling issues (Mercer and Wainwright, 2008). Many of the apparent inconsistencies among these studies are most likely due to insufficient evaluation of estimation errors. Establishing the statistical uncertainty of gene flow estimates is as important as minimizing their bias for incorporating scientific advice into decision making. However, this aspect of estimation is frequently neglected.

We consider here the illustrative example of Scots pine (*Pinus sylvestris* L.) populations in Mediterranean Spain. *Pinus sylvestris* var. *nevadensis* D. H. Christ is one of the

Correspondence: Dr JJ Robledo-Arnuncio, Departamento de Sistemas y Recursos Forestales, CIFOR-INIA, Unidad de Genética y Ecofisiología Forestal, Ctra. de la Coruña km 7.5, Madrid 28040, Spain.
E-mail: jjrobledo@gmail.com
Received 12 January 2009; revised 5 May 2009; accepted 15 May 2009; published online 15 July 2009

most widely recognized varieties (Farjon, 1998) of this monoecious, wind-pollinated, predominantly outcrossing conifer. It is restricted to a few relict populations in southern Spain, comprising the Trevenque population, which is scattered over 120 hectares at the southernmost extremity of the pine's range, within the Sierra Nevada National Park. This area is closely surrounded by several thousand hectares of conspecific tree plantations not belonging to the *nevadensis* variety (Figure 1). Introgression from the plantations is a current matter of concern for the National Park managers and local foresters, because mtDNA (Sinclair *et al*., 1999), cpDNA (Provan *et al*., 1998) and quantitative growth and survival traits (Alía *et al*., 2001) have shown a substantial genetic divergence of *P. sylvestris* var. *nevadensis* relative to other Iberian and European populations of the species.

Few formal procedures are available for estimating contemporary seed and pollen migration over broad spatial scales, as required to assess genetic introgression into the Scots pine relict. Genetic parentage exclusion methods have been developed for assessing immigration into small populations, such as seed orchards (Adams *et al*., 1997; Plomion *et al*., 2001; Stoehr and Newton, 2002; Slavov *et al*., 2005), but these methods are not feasible for large-scale experiments, because they require exhaustive genotyping of all potential parents within the recipient population. Genetic assignment methods, on the other hand, have been designed primarily to ascertain population membership of individuals and admixture proportions, and are not suitable for addressing the statistical question of obtaining accurate migration rates exclusively (Manel *et al*., 2005). Although they can be used to obtain rough migration rate estimates (Manel *et al*., 2005), assignment methods typically offer low power when candidate populations are genetically close and do not account for cryptic immigration (immigrants wrongly identified as non-immigrants) when deriving migration rate estimates (Cornuet *et al*., 1999; Paetkau *et al*., 2004).

The Bayesian approaches of Wilson and Rannala (2003) and Faubet and Gaggiotti (2008) estimate posterior probability distributions of several population parameters, including recent migration rates among populations. Although these methods are potentially useful for conservation plans requiring recent genetic immigration measures, their accuracy and the reliability of the associated credibility intervals have not been tested under varied realistic demographic conditions. This is not an easy task to perform numerically, given the slowness and frequent convergence problems of the Bayesian estimation algorithms (Faubet *et al*., 2007; Faubet and Gaggiotti, 2008), which limit their utility within a risk assessment context. In addition, these advanced procedures use diploid biparentally inherited markers and are designed to address the general problem of jointly estimating gene migration rates, population frequencies, inbreeding coefficients and the origin of every individual in the sample. We lack more specific and efficient approaches to estimate migration rates exclusively, for which we could exploit the methodological and inferential advantages of haploid uniparentally inherited DNA markers.

Here we use a straightforward maximum-likelihood (ML) procedure to estimate recent male gametic immigration rates based on paternally inherited chloroplast DNA markers. We assume a single external source of potential immigrants, albeit extension to more than one source population is straightforward. Unlike parentage exclusion methods, our approach does not require exhaustive genotyping of the recipient population. As opposed to assignment methods (Paetkau *et al*., 1995, 2004; Rannala and Mountain, 1997; Cornuet *et al*., 1999; Pella and Masuda, 2001; Wilson and Rannala, 2003; Faubet and Gaggiotti, 2008), here we do not aim at establishing population membership of every individual in the sample, but only to obtain an accurate immigration rate estimate and place reliable confidence limits around it. We apply this method to estimate allochthonous male gametic immigration into the Scots pine relict population in the Sierra Nevada National Park, showing how researchers can improve decision making within a gene flow risk assessment context.

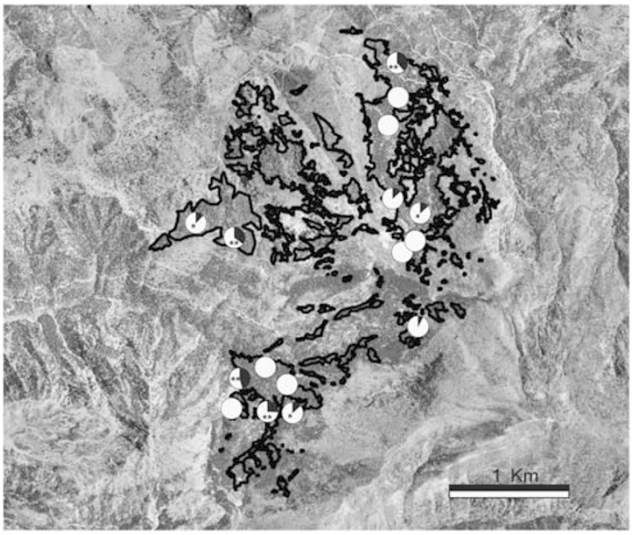

Figure 1 Male gametic introgression rates (in black) from non-local plantations into the natural regeneration plots of *Pinus sylvestris* L. var. *nevadensis* relict population at Trevenque (Sierra Nevada, Spain). Lines delimit the natural stands, with all the observable surrounding forests being non-local conspecific plantations. The seven southernmost plots are in a dense area (REG1, named 'La Cortijuela'), whereas the remaining nine are within sparser stands growing on a different slope (REG2, named 'La Dehesilla'). The estimates significantly different from 0 (95% confidence interval not containing 0) are indicated by * (if only direct estimation of haplotype frequencies yielded a positive test) or by ** (if both direct and Bayesian estimation of haplotype frequencies yielded positive tests).

Materials and methods

Field sampling

Sampling was conducted in three stages. We first collected needle tissue from $n_N = 112$ and $n_P = 108$ adult individuals from the native *Pinus sylvestris* var. *nevadensis* population (N) and the allochthonous plantations (P), respectively, which we used to select discriminating markers and estimate adult haplotypic frequencies. Adult trees from N were more than a century old and were naturally established well before P was introduced about 50 years ago. We distinguished two sampling areas within the natural population: a dense area (N1, locally

known as 'La Cortijuela'; $n_{N1}=63$) and a sparser one on a different slope (N2, locally known as 'La Dehesilla'; $n_{N2}=49$) (Figure 1). We extracted total DNA and screened all 220 adult trees at six chloroplast microsatellite markers (cpSSR)—Pt1254, Pt15169, Pt26081, Pt30204, Pt36480 and Pt87268 (Vendramin *et al*., 1996)—according to the procedures detailed by Robledo-Arnuncio *et al*. (2005). After the initial screening, we identified four markers (Pt15169, Pt26081, Pt36480 and Pt87268) that provided most of the discriminating power between adult individuals of N and P, including some discriminant size variants that were present only in the plantations.

Second, in spring 2004, we sampled a total of 325 recruits (0–30 years old) from 16 natural regeneration plots distributed under the canopy of N (Figure 1). We used this sample to estimate the male gametic introgression rate (m) at the level of regeneration, which may occur either through seed dispersal from N mothers that have been pollinated by P fathers or by seed dispersal from P mothers pollinated by P fathers. Reproductive recruits were very rare and we considered that no backcrosses had taken place in the population. Our sampling estimated the accumulated proportion of immigrant male gametes arriving from the plantations from the time planted trees reached maturity (about 30 years ago) to the present. Seven of the plots ($n_S=138$ individuals) were in the dense N1 area (REG-1), and the other nine plots ($n_S=187$ individuals) were within the sparser N2 area (REG-2). The mean number of sampled recruits per plot was 20.3 (s.d. = 6.0). Each of the 325 samples from the natural regeneration was scored at the same four cpSSR that adult individuals were scored at.

Finally, in winter 2004, we collected a total of $n_S=440$ seeds from 22 seed trees at 11 sampling locations within the REG-1 area (one pair at each location), with a fixed number of 40 seeds per tree pair (Figure 2). We used this sample to estimate male gametic introgression (possible only through pollen in this case) from P into the standing seed crop of N over the pollen dispersal episode of the year 2003. Seeds were germinated and the embryo of each seed was characterized with the selected cpSSR.

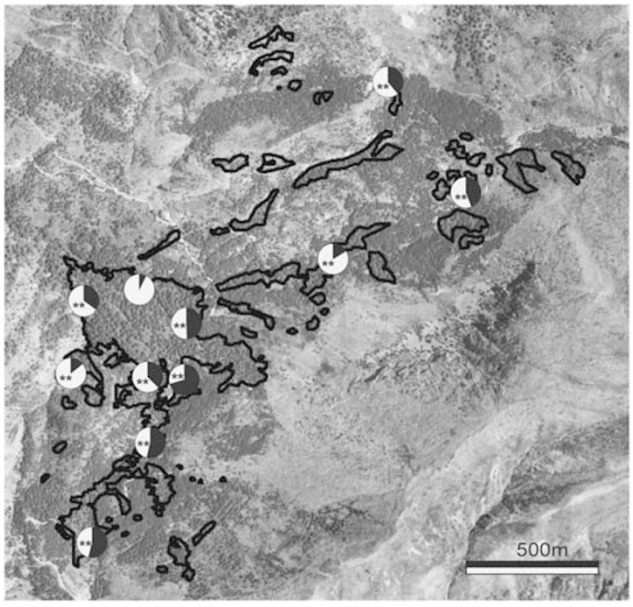

Figure 2 Pollen immigration rates (in black) from non-local plantations into the standing seed crop of Scots pine relict population at 'La Cortijuela', Trevenque. The estimates significantly different from 0 (95% confidence interval not containing 0) are indicated by * (if only direct estimation of haplotype frequencies yielded a positive test) or by ** (if both direct and Bayesian estimation of haplotype frequencies yielded positive tests).

Model for estimating immigration rates

We defined recent male gametic immigration rate (m) as the proportion of male gametes arriving at N from P during the reference migration period. In our study at Trevenque, the reference period was the previous 30 years (since the plantations reached maturity) for naturally regenerating seedlings and the 2003 pollination episode for the seed crop collected in 2004. Consider a sample S of n_S individuals from population N, collected after the reference migration period, for which we want to know m. This would be either the sample of seeds or naturally regenerating seedlings. Let k be the number of chloroplast haplotypes (cp-haplotypes), defined as unique combinations of variants at the set of chloroplast DNA marker regions, and let $q_{h,N}$ and $q_{h,P}$ be the frequencies before the reference migration episode of the h-th haplotype in populations N and P, respectively. Under random mating, the probability that cp-haplotype h is observed in the sample S (after migration) is

$$\Pr(h|m) = (1-m)q_{h,N} + mq_{h,P} \quad (1)$$

In the favourable case that there was a set of l plantation-specific cp-haplotypes ($h=1, 2,\ldots, l$ with $q_{h,N}=0$), a direct estimator of the male gametic immigration rate would then be

$$\hat{m} = \sum_{h=1}^{l} q_{h,S} \Big/ \sum_{h=1}^{l} q_{h,P} \quad (2)$$

where $q_{h,S}$ is the frequency of the h-th cp-haplotype in the sample S. More generally, we can use equation (1) to compute the joint-likelihood function for the whole set of cp-haplotypes (that is, including both plantation-specific and shared ones) carried by the n_S individuals in the sample S as a function of m:

$$L(m|S) = \prod_{h=1}^{n_S} \Pr(h|m) \quad (3)$$

and estimate m by maximizing the log-likelihood function

$$\ln L(m|S) = \sum_{h=1}^{n_S} \ln[(1-m)q_{h,N} + mq_{h,P}] \quad (4)$$

Note that the estimator of m obtained from equation (4) accounts for cryptic immigration, as it defines the probability of observing a cp-haplotype in the recipient population as a function of m and its frequency in both the candidate source ($q_{h,P}$) and recipient ($q_{h,N}$) populations. On the basis of the estimated adult population cp-haplotype frequencies (see section 'Estimating population haplotype frequencies'), we used either equation (2) or (4) to obtain two estimates of m at each of the two natural regeneration areas REG-1 and REG-2 ($n_S=138$ and 187 seedling cp-haplotypes, respectively) and at the seed-crop level ($n_S=440$ seed embryo cp-haplotypes). In order to show potential spatial heterogeneities in

introgression levels, we also estimated m separately at each of the 16 natural regeneration plots and for each of the 40 embryo samples collected from each seed-tree pair.

Estimating population haplotype frequencies

The population cp-haplotype frequencies before migration (the $q_{h,\mathrm{N}}$−s and the $q_{h,\mathrm{P}}$−s) must be estimated from the field samples. Let $x_{h,\mathrm{N}}$ and $x_{h,\mathrm{P}}$ be the observed counts of the h-th cp-haplotype in the adult samples from populations N and P, respectively (with $h=1, 2,\ldots, k$). Maximum-likelihood cp-haplotype frequency estimates are given by

$$\begin{aligned}\hat{q}_{h,\mathrm{N}} &= x_{h,\mathrm{N}}/n_{\mathrm{N}} \\ \hat{q}_{h,\mathrm{P}} &= x_{h,\mathrm{P}}/n_{\mathrm{P}}\end{aligned} \tag{5}$$

If the sample sizes n_{N} and n_{P} were very large, we might safely assume that the cp-haplotype frequencies estimated using equation (5) are close enough to the exact population values, and use them in equation (4) to obtain an immigration estimate $\hat{m}$. In a decision-making context, however, it may be appropriate to minimize the type I error (non-immigrants wrongly identified as immigrants) in order to avoid potential overestimates of m, which could motivate unnecessary and costly actions to reduce introgression. A source of type I error could be the presence of low-frequency plantation-specific cp-haplotypes in the observed sample ($\hat{q}_{h,\mathrm{N}}=0$ and a small $\hat{q}_{h,\mathrm{P}}>0$), if in fact they are present (but not detected) in the natural population (small $q_{h,\mathrm{N}}>0$). We could then assume *a priori* that all cp-haplotypes are shared between P and N at equal frequency, and estimate haplotype population frequencies on the basis of this prior assumption and the observed counts. For this purpose, we use the Bayesian approach described by Rannala and Mountain (1997), which gives posterior cp-haplotype frequencies as

$$\begin{aligned}\hat{q}'_{h,\mathrm{N}} &= \frac{x_{h,\mathrm{N}} + 1/K}{n_{\mathrm{N}} + 1} \\ \hat{q}'_{h,\mathrm{P}} &= \frac{x_{h,\mathrm{P}} + 1/K}{n_{\mathrm{P}} + 1}\end{aligned} \tag{6}$$

where K is the total number of cp-haplotypes in the two populations, which we estimate with the number (k) of observed cp-haplotypes in the sample. By using equation (4) and the cp-haplotype frequencies estimated with equation (6), we obtain a conservative immigration rate estimate ($\hat{m}'$) accounting for potential overestimates derived from undetected low-frequency cp-haplotypes in the recipient population.

Assessing statistical uncertainty

Given the two vectors of estimated cp-haplotype frequencies for populations N $\{\hat{q}_{1,\mathrm{N}}, \hat{q}_{2,\mathrm{N}}, \ldots, \hat{q}_{k,\mathrm{N}}\}$ and P $\{\hat{q}_{1,\mathrm{P}}, \hat{q}_{2,\mathrm{P}}, \ldots, \hat{q}_{k,\mathrm{P}}\}$, we computed the expected bias and accuracy (root mean square error, RMSE) of $\hat{m}$ for a sample size n_{S} and an assumed introgression rate m by simulating stochastic samples of size n_{S} as follows: (i) draw a random number r in the half-closed interval [0, 1); (ii) generate an individual i with assigned source population P if $r<m$, or with assigned source population N otherwise; (iii) randomly draw a cp-haplotype for individual i from a multinomial distribution with k possible outcomes, with respective probabilities $\{\hat{q}_{1,\mathrm{N}}, \hat{q}_{2,\mathrm{N}}, \ldots, \hat{q}_{k,\mathrm{N}}\}$ if i comes from N, or with respective probabilities $\{\hat{q}_{1,\mathrm{P}}, \hat{q}_{2,\mathrm{P}}, \ldots, \hat{q}_{k,\mathrm{P}}\}$ if i comes from P; (iv) start again from the first step until obtaining n_{S} individuals. Next, we compute $\hat{m}$ by applying equation (2) (or equation (4)) to the simulated sample of size n_{S} and repeat the whole procedure to generate 10 000 independent random samples with their corresponding $\hat{m}$ values. On the basis of the 10 000 $\hat{m}$ values, we compute the expected relative bias and expected relative RMSE of the m estimator for the assumed sample size, migration rate and cp-haplotype frequencies.

Next, we assessed the performance of four common methods to estimate confidence intervals (CIs): the standard bootstrap (S-Boot), the simple percentile bootstrap (SP-Boot), the accelerated bias-corrected percentile bootstrap (ABC-Boot) and the profile-likelihood method (P-Likelihood). S-Boot is based on variance estimates and the assumption of normality for the distribution of the estimator, whereas SP-Boot is free of this assumption. ABC-Boot accounts for potential biases arising from skewed bootstrap distributions. Finally, P-Likelihood is based on the inspection of the profile-likelihood function. More detailed descriptions of these statistical procedures can be found elsewhere (for example, Manly, 1997, pp 34–55; Coles, 2001, pp 34–35). On the basis of the simulated data obtained with the Monte Carlo approach described above, we estimated 95% CIs for $\hat{m}$, one for each of the 10 000 independent samples, using these four different methods. Next, we computed the coverage of each method as the proportion of the 10 000 estimated CIs that contained the assumed immigration estimate (m) in the simulations (Manly, 1997). Specifically, we assessed the proportion of times that the upper confidence limit was too low (smaller than m) and the proportion of times that the lower confidence limit was too high (larger than m), comparing these proportions to their nominal 2.5% value.

In the simulations, we assumed that haplotype population frequencies were equal to the ML frequencies (equation (5)), and we considered four sample sizes: $n_{\mathrm{S}}=20$, 40, 138 and 440, corresponding to the real sample sizes used in our empirical study. For each sample size, we assumed migration rates (m) ranging from 0 to 100%. We did not intend to evaluate the statistical behaviour of the methods under a broad range of sampling, genetic differentiation and immigration conditions, but rather to exemplify the assessment of the statistical uncertainty of introgression estimates for a particular data set in a real decision-making scenario.

Results

Genetic diversity and haplotype frequencies

Haplotype richness among adult trees was substantially lower in the Scots pine relict population (N) than in the surrounding plantations (P), with 7 and 24 observed cp-haplotypes, respectively (Table 1). The effective number of cp-haplotypes, calculated as the inverse of the unbiased haplotypic diversity, was 2.8 in N and 7.4 in P. Haplotypic differentiation between N (N1 and N2 pooled together) and P was $F_{\mathrm{ST}}=0.048$ ($P<0.01$), whereas that between N1 and N2 was $F_{\mathrm{ST}}=0.034$ ($P=0.041$). Although all cp-haplotypes found in N were present in P, 17 putative plantation-specific cp-haplotypes

Table 1 Estimated chloroplast haplotype frequencies among Scots pine adult trees, naturally regenerated recruits and seeds at Trevenque

Haplotype		*Adults native population* (N)				*Adults plantation* (P)		*Regeneration*		*Seed crop (from* N1)
Code	*Size variants*[a]	N1		N2		$\hat{q}_{h,\mathrm{P}}$	$\hat{q}'_{h,\mathrm{P}}$	*REG-1*	*REG-2*	
		$\hat{q}_{h,\mathrm{N1}}$	$\hat{q}'_{h,\mathrm{N2}}$	$\hat{q}_{h,\mathrm{N1}}$	$\hat{q}'_{h,\mathrm{N2}}$					
H1	165-144-110-126	0.5079	0.5007	0.6327	0.6208	0.3056	0.3031	0.4928	0.7540	0.4045
H2	166-144-110-125	0.2063	0.2038	0.0612	0.0608	0.0833	0.0830	0.1449	0.0802	0.1114
H3	165-144-110-127	0.0159	0.0163	0.0204	0.0208	0.0926	0.0921	0.0362	0.0267	0.0591
H4	165-144-109-126	0.0476	0.0475	0.0816	0.0808	0.0093	0.0096	0.0362	0.0374	0.0227
H5	165-144-110-125	0.2063	0.2038	0.0612	0.0608	0.1481	0.1472	0.2464	0.0214	0.2341
H6	165-144-109-125	0.0159	0.0163	0.1020	0.1008	0.0093	0.0096	0	0.0053	0.0068
H7	165-143-110-126	0	0.0007	0.0408	0.0408	0.0741	0.0738	0.0145	0.0374	0.0409
H8	165-143-110-124	0	0.0007	0	0.0008	0.0093	0.0096	0	0	0
H9	165-144-110-124	0	0.0007	0	0.0008	0.0463	0.0463	0	0	0.0045
H10	165-143-111-126	0	0.0007	0	0.0008	0.0185	0.0187	0	0	0.0023
H11	165-144-111-126	0	0.0007	0	0.0008	0.0185	0.0187	0	0	0.0045
H12	167-144-110-125	0	0.0007	0	0.0008	0.0093	0.0096	0	0	0.0023
H13	166-144-110-127	0	0.0007	0	0.0008	0.0093	0.0096	0	0	0.0068
H14	165-144-109-128	0	0.0007	0	0.0008	0.0278	0.0279	0	0	0.0114
H15	168-144-110-125	0	0.0007	0	0.0008	0.0278	0.0279	0	0	0.0045
H16	166-144-110-128	0	0.0007	0	0.0008	0.0093	0.0096	0	0	0.0023
H17	165-143-111-125	0	0.0007	0	0.0008	0.0185	0.0187	0.0145	0	0.0023
H18	168-144-111-126	0	0.0007	0	0.0008	0.0185	0.0187	0	0	0.0045
H19	165-143-110-125	0	0.0007	0	0.0008	0.0093	0.0096	0.0072	0	0.0114
H20	166-144-110-124	0	0.0007	0	0.0008	0.0185	0.0187	0	0	0.0023
H21	164-144-109-125	0	0.0007	0	0.0008	0.0093	0.0096	0	0	0
H22	165-144-112-125	0	0.0007	0	0.0008	0.0093	0.0096	0	0	0
H23	165-144-110-128	0	0.0007	0	0.0008	0.0093	0.0096	0	0	0
H24	166-144-109-125	0	0.0007	0	0.0008	0.0093	0.0096	0.0072	0.0374	0.0091

Two estimators of haplotype frequency were used for adult trees: maximum-likelihood ($\hat{q}_h$, using equation (5)) and Bayesian ($\hat{q}'_h$, using equation (6)), the latter under the earlier assumption that all haplotypes are shared between the native population and the plantation.
[a]Size variants from Pt87268–Pt36480–Pt26081–Pt15169 chloroplast markers.

were detected at low individual ($\hat{q}_{h,\mathrm{P}} = 0.009$–$0.0463$) but substantial cumulative ($\sum \hat{q}_{h,\mathrm{P}} = 0.278$) frequencies (Table 1). Under the prior assumption that all haplotypes are shared between N and P, individual posterior frequencies in the natural stands for putatively plantation-specific cp-haplotypes were low: $\hat{q}'_{h,\mathrm{N1}} = 0.0007$ for N1 and $\hat{q}'_{h,\mathrm{N2}} = 0.0008$ for N2 (Table 1).

Most of the 325 naturally regenerated recruits collected below the canopy of N showed the same set of seven cp-haplotypes shared among the adults of N and P, but a few of the putatively plantation-specific cp-haplotypes were also detected at low frequency among the recruits (Table 1). The observed cumulative frequencies of putatively plantation-specific cp-haplotypes were 0.029 at REG-1 and 0.0374 at REG-2. Among the 440 seeds collected from the standing crop of adult trees of N, a total of 387 carried cp-haplotypes shared between adult trees from N and P (Table 1), whereas 30 seeds carried putative plantation-specific cp-haplotypes (with a cumulative frequency of 0.068; Table 1), and the remaining 23 seeds showed 23 different cp-haplotypes that were absent among sampled adult trees of both N and P (data not shown).

The 23 rare haplotypes (frequency = 0.0023) found in the seed crop were probably present in the adult populations N and P but remained undetected because of the smaller adult sample sizes. An alternative, yet unlikely, explanation is that some haplotypes might have dispersed from distant populations (the closest conspecific population is more than 100 km away) or may have arisen by mutation or genotyping error. In either case, as the estimated frequencies for these 23 haplotypes are identical among adult trees of N and P (0 if estimated directly or a small value close to 0 if estimated using the Bayesian approach), they lack inferential value for the estimation of male gametic immigration from P into N, so we discarded them from subsequent analysis.

Immigration estimates

The estimates of male gametic immigration (m) from P into N did not differ significantly when considering adult tree frequencies of N1 and N2 pooled together or separately (results not shown), so we only report here estimates based on the latter. When only putatively plantation-specific cp-haplotypes were used (equation (2)), estimated immigration rates of male gametes from the plantation into the natural population were $\hat{m} = 0.029/0.2778 = 0.104$ for REG-1, $\hat{m} = 0.0374/0.2778 = 0.135$ for REG-2 and $\hat{m} = 0.068/0.2778 = 0.245$ for the seed crop. On the basis of both plantation-specific and shared cp-haplotypes (equation (4)), estimates of male gamete immigration when directly estimating adult haplotype frequencies with equation (5) were $\hat{m} = 0.143$ (REG-1), 0.121 (REG-2) and 0.387 (seed crop), whereas the corresponding values under the prior assumption that all cp-haplotypes are shared between the natural population and the plantation (equation (6)) were $\hat{m} = 0.116$ (REG-1), 0.044 (REG-2) and 0.369 (seed crop) (Table 2).

Immigration estimates at the regeneration-plot level ranged from 0 to 45%, and did not reveal a clear spatial pattern; plots showing larger values appear scattered over the whole area (Figure 1). Pollen immigration at the seed-crop level ranged from 7 to 70% among seed trees, and in this case, progenies showing the highest

immigration rates were generally those from seed trees located in narrow elongated areas or small fragments embedded within the plantations (Figure 2).

Expected bias and accuracy

Simulation results indicated that m-estimates exclusively based on plantation-specific cp-haplotypes (equation (2)) had as low a bias as ML estimates based on the full set of observed haplotypes (equation (4)), but were consistently less accurate (relative RMSE 5–80% larger) for all sample sizes and assumed immigration rates (results not shown). The sensitivities of the two estimators to both sample size and assumed immigration rate were qualitatively and quantitatively very similar. Therefore, here we comment only on the results corresponding to the more accurate ML approach.

The ML estimation method (equation (4)) is expected to yield male gametic immigration estimates with very low bias for all n_S and m values considered (Table 3). The expected relative bias was smaller than 6% in all cases, and smaller than 3% in most cases and did not show a constant sign across simulation conditions. The expected relative RMSE was sensitive to both n_S and m, with a marked decreasing trend with increasing n_S and m. This trend was determined by the higher variance of the estimator for small sample sizes and small parametric m values ($MSE = Bias^2 + Variance$). The relative RMSE reached a high of 1.8 for $n_S = 20$ and $m = 0.05$, whereas it remained below 0.4 for any value of m when $n_S = 440$ (Table 3). Considering the real sample sizes and estimated immigration rates for Scots pine, the ML estimator is expected to be virtually unbiased in all cases, while its expected relative RMSE would be about 0.4 for REG-1 ($n_S = 138$, $\hat{m} = 0.143$), 0.35 for REG-2 ($n_S = 187$, $\hat{m} = 0.121$) and 0.1 for the seed crop ($n_S = 440$, $\hat{m} = 0.387$). The expected relative RMSE for immigration estimates at the regeneration plot ($n_S = 20$) and seed-tree pair ($n_S = 40$) levels are more variable and generally higher, ranging from 0.2 to 1.8 for different values of m (Table 3).

Table 2 Estimates of the proportion of immigrant male gametes from non-local plantations into the natural regeneration areas (REG-1 and REG-2) and the standing seed crop of the Scots pine relict population at Trevenque

	ML-frequency		*Bayesian-frequency*	
	$\hat{m}$	*(95% CI)*	$\hat{m}'$	*(95% CI)*
REG-1	0.143	(0.061–0.267)	0.116	(0.033–0.243)
REG-2	0.121	(0.055–0.217)	0.044	(0.000–0.151)
Seed crop	0.387	(0.300–0.480)	0.369	(0.280–0.462)

Immigration estimates are based on either maximum-likelihood ($\hat{m}$) or Bayesian ($\hat{m}'$) estimates of haplotype frequencies (equations (5) and (6), respectively). 95% confidence intervals (CIs) were computed using the profile-likelihood method (for REG-1 and REG-2) and simple-percentile bootstrapping (for seed crop).

Confidence interval estimation

Simulations indicate that the four tested methods for finding CIs for m-estimates yield considerably different coverage for constant 95% nominal limits, and that all methods generally perform better for larger n_S and m (Table 4). Overall, P-Likelihood clearly gave the best and most robust performance, with its limits including the parametric m between 91 and 100% of the time across the range of assumed n_S and m values. Both the upper and lower limits calculated by P-Likelihood were quite satisfactory, ranging from 0 to 5%, close to the nominal 2.5%. By contrast, the three bootstrapping approaches gave generally bad performances and were far more sensitive to n_S and m. For $n_S < 138$ and $m < 0.1$, all S-Boot, SP-Boot and ABC-Boot methods yielded CIs including the parametric m as infrequently as 65% of the time, with their upper limit being too low 20–30% of the time. For larger sample size ($n_S = 440$), all four methods tended to give similar performances, especially when $m > 0.2$, with SP-Boot very slightly outperforming P-Likelihood for $m = 0.4$ (Table 4).

On the basis of the simulation results, we chose P-Likelihood for estimating CIs for field estimates of m at the regeneration-plot and seed-tree pair levels, as well as for the global natural regeneration estimates. We restricted the use of SP-Boot for estimating the CI of $\hat{m}$ for the global seed-crop sample. The results indicate that the global male gametic immigration estimates are significantly different from 0 both at the regeneration (REG-1 and REG-2) and seed-crop level when using ML estimates of cp-haplotype frequencies (Table 2). However, considering the conservative prior assumption that the natural population and the plantation share all cp-haplotypes, we cannot reject the hypothesis that the global immigration rate is 0 at REG-2 (Table 2). On a finer scale, only the largest estimates were significantly different from 0 at the regeneration-plot level, but

Table 3 Expected relative bias (Bias) and expected relative root mean square error (RMSE) of the maximum-likelihood estimator of the proportion of immigrant male gametes based on the full set of observed haplotypes (using equation (4))

m	$n = 20$		$n = 40$		$n = 138$		$n = 187$		$n = 440$	
	Bias	*RMSE*	*Bias*	*RMSE*	*Bias*	*RMSE*	*Bias*	*RMSE*	*Bias*	*RMSE*
0	0.006	0.031	0.001	0.010	0.000	0.000	0.000	0.000	0.000	0.000
0.05	0.032	1.775	−0.001	1.288	−0.018	0.692	−0.002	0.599	−0.003	0.382
0.10	−0.023	1.218	−0.013	0.885	0.001	0.469	0.002	0.399	0.005	0.259
0.15	−0.015	0.983	−0.003	0.703	0.001	0.366	0.000	0.314	0.001	0.204
0.20	−0.021	0.828	0.005	0.591	−0.003	0.308	0.001	0.264	0.002	0.173
0.40	0.000	0.533	0.003	0.371	0.000	0.193	−0.001	0.168	0.000	0.108
0.80	0.018	0.335	0.010	0.222	0.000	0.110	−0.002	0.094	0.000	0.062
1	0.055	0.296	0.033	0.200	0.005	0.089	0.002	0.072	0.001	0.045

Error estimates are based on 10 000 independent simulations of the immigration process, assuming different sample sizes (n) and immigration rates (m). Errors for $m = 0$ are absolute, as relative values are undefined in this case.

Table 4 Coverage probabilities, for a nominal 95% level, of four methods to estimate confidence intervals for male gametic immigration estimates: standard bootstrap (S-Boot), simple percentile bootstrap (SP-Boot), accelerated bias-corrected percentile bootstrap (ABC-Boot) and profile-likelihood (P-Likelihood) method (Manly, 1997; Coles, 2001)

n	*m*	*S-Boot*		*SP-Boot*		*ABC-Boot*		*P-Likelihood*	
		LLTH	*ULTL*	*LLTH*	*ULTL*	*LLTH*	*ULTL*	*LLTH*	*ULTL*
20	0	0.02	0	0.02	0	0.32	0	0.04	0
	0.05	0.05	*33.49*	0.04	*32.60*	0.92	*31.99*	2.03	0
	0.10	0.19	*33.63*	0.40	*23.61*	*2.98*	*20.42*	1.97	0
	0.15	0.51	*28.02*	0.77	*19.19*	*3.20*	*12.85*	2.35	0
	0.20	0.55	*22.30*	0.82	*15.91*	*3.17*	*8.46*	2.12	1.90
	0.40	1.00	*9.69*	1.69	*6.89*	*3.25*	*3.98*	*2.60*	*4.24*
	0.80	*2.84*	*4.08*	*3.63*	*3.35*	*5.34*	*3.35*	*3.03*	*2.82*
	1	1.17	2.22	*5.89*	1.92	*7.26*	*2.95*	0	2.37
40	0	0	0	0	0	*34.95*	0	0	0
	0.05	0.08	*37.76*	0.22	*31.69*	*7.77*	*16.78*	2.15	0
	0.10	0.45	*25.30*	0.63	*19.73*	*4.00*	*11.20*	2.23	0.36
	0.15	0.54	*15.54*	0.84	*11.69*	*4.26*	*6.30*	2.38	*5.28*
	0.20	0.93	*11.12*	1.24	*7.73*	*4.26*	*3.58*	2.48	*4.72*
	0.40	1.72	*5.74*	2.08	*4.56*	*4.38*	*3.02*	*2.61*	*2.90*
	0.80	*2.73*	*2.98*	*3.70*	*2.61*	*4.27*	*4.18*	*3.20*	2.38
	1	0	1.55	0	1.54	0	*3.50*	0	2.09
138	0	0	0	0	0	*4.44*	*3.81*	2.31	*3.02*
	0.05	0.32	*14.85*	0.63	*14.27*	*4.67*	*4.26*	2.38	*2.74*
	0.10	0.76	*8.55*	1.10	*6.70*	*4.69*	*4.59*	2.37	2.36
	0.15	1.00	*6.49*	1.34	*5.12*	*4.20*	*4.82*	*2.60*	2.36
	0.20	1.28	*5.45*	1.52	*4.44*	0	*4.57*	0	2.22
	0.40	1.78	*3.68*	1.97	*3.35*	0.23	0	0	0
	0.80	*2.52*	*2.52*	*2.83*	*2.52*	*4.65*	*3.65*	2.19	*2.72*
	1	0	1.69	0	1.82	*4.94*	*4.25*	2.48	2.25
440	0	0	0	0	0	*4.95*	*4.89*	2.20	2.43
	0.05	0.69	*7.39*	1.02	*5.65*	*4.95*	*4.92*	2.18	2.48
	0.10	1.30	*4.60*	1.64	*3.93*	*4.72*	*4.77*	2.32	2.13
	0.15	1.40	*4.47*	1.62	*3.80*	*4.78*	*5.30*	*2.57*	*2.76*
	0.20	1.56	*3.82*	1.74	*3.46*	0	*4.95*	0	2.30
	0.40	2.00	*2.93*	2.17	*2.66*	*4.44*	*3.81*	2.31	*3.02*
	0.80	*2.81*	*2.88*	*2.63*	*2.98*	*4.67*	*4.26*	2.38	*2.74*
	1	0	1.95	0	2.16	*4.69*	*4.59*	2.37	2.36

The percentage of times that the lower confidence limit was too high (LLTH) and the upper limit too low (ULTL) were assessed based on 10 000 independent simulations of the immigration process, assuming different sample sizes (*n*) and immigration rates (*m*). Values exceeding nominal limits are shown in italics.

virtually all estimates were significantly different from 0 at the seed-tree pair level (Figures 1 and 2).

Discussion

We used a novel and straightforward ML method, on the basis of haploid uniparentally inherited genetic markers, to estimate recent uniparental gametic immigration from non-local plantations into a relict Scots pine population. By exclusively estimating immigration rates, we circumvented convergence issues, large computation time and potential biases associated with more general methods involving the joint estimation of migration rates, population membership of individuals, allelic frequencies and other population coefficients (Pella and Masuda, 2001; Wilson and Rannala, 2003; Faubet *et al*., 2007; Faubet and Gaggiotti, 2008). The Monte Carlo simulation results show that the proposed method can be expected to yield virtually unbiased and fairly accurate uniparental gametic immigration estimates under our sampling conditions. In addition, we have illustrated how our approach fits within a gene flow risk assessment framework, allowing a reliable evaluation of statistical uncertainty of immigration estimates.

Our definition of recent migration refers to the exchange of migrants occurred during a reference migration period, such as episodes of pollen and seed dispersal, before and after which we can unambiguously estimate haplotype population frequencies. This sequential sampling approach, feasible for organisms with discrete synchronized migration periods, avoids the necessary joint estimation of allele frequencies (before migration) of methods that compute migration rates using the target sample of potential immigrants also as reference set (Wilson and Rannala, 2003; Faubet and Gaggiotti, 2008).

The estimation of uniparental gametic immigration using haploid uniparentally inherited markers represents a more tractable problem than obtaining total immigration rates from diploid biparentally inherited genetic data. We neither required assuming Hardy–Weinberg proportions or linkage equilibrium among loci nor required jointly estimating inbreeding coefficients. Consequently, the simplicity of the log-likelihood function (equation (4)) allows an extremely fast optimization that facilitates the evaluation of the robustness and accuracy of the estimator. As pointed out by Wilson and Rannala (2003), it is advisable to evaluate numerically the reliability of migration rate estimates for particular data sets, given the observed levels of genetic differentiation. However, assessing the statistical properties of an estimator on the basis of as few as 10 replicates, a frequently adopted limit imposed by the time-consuming available Bayesian methods to estimate recent migration (Wilson and Rannala, 2003; Faubet *et al*., 2007; Faubet and Gaggiotti, 2008) may be insufficient (Manly, 1997). This limitation can be a serious drawback in real risk-assessment scenarios. By contrast, we were able to conduct heavily (10 000) replicated simulations in our case study, which provides a sounder basis for calculating the expected accuracy of migration estimates and choosing a reliable method for deriving CIs. Indeed, as suggested by our simulation results, uninformed method selection for assessing statistical uncertainty could yield flawed and misleading CIs for migration rates.

Our numerical analyses exemplify the assessment of the statistical uncertainty of introgression estimates for a particular data set in a real decision-making scenario. Future theoretical studies should evaluate the statistical behaviour and potential limitations of the proposed estimation method under a broad range of sampling, demographic and genetic differentiation conditions. Of particular interest will be how the accuracy of immigration estimates and uncertainty measures are affected by very low levels of genetic differentiation ($F_{ST} \ll 0.05$) among immigrant sources and/or by the presence of multiple (sampled or unsampled) potential source populations.

The possibility of obtaining uniparental gametic immigration measures on a landscape scale is relevant to many scientific and practical problems, such as evolutionary investigations about the consequences of sex-biased dispersal among populations (Hu and Ennos, 1999; Lopez *et al*., 2008) or conservation studies evaluating the relative exposure of natural plant populations to seed and pollen from exotic or transgenic plantations.

When using paternally inherited markers, as in the Scots pine case study, migration rate estimates will measure pollen immigration only if the target sample is a set of seeds collected before dispersal, whereas estimates will reflect immigration of male gametes by both pollen and seed dispersal if the sample is a set of individuals collected after seed dispersal. Alternatively, given that the only possible vector for female gametic immigration is seed dispersal, estimates obtained using haploid maternally inherited markers would necessarily measure seed immigration rates (note that our approach could also be used for this purpose). By combining genetic markers with different modes of uniparental inheritance, it would be possible to estimate ratios of male to female gamete immigration rates.

The empirical results from this study reveal some considerations for conservation management of the Scots pine relict population at Trevenque. Our estimates indicate very high (about 40%) pollen introgression from non-local plantations at the seed-crop level before seed dispersal and substantial (10–15%) male gametic introgression among naturally regenerated recruits, both being significantly different from 0. If managers chose to work under the conservative prior assumption that all haplotypes are shared between the populations, however, we could not conclude that the introgression levels are significantly different from 0 among recruits sampled in the 'La Dehesilla' area (REG-2), which, based on this evidence, could be considered a candidate for a gene reserve.

Interestingly, male gametic immigration was about threefold higher at the seed-crop level than at the regeneration level ($P = 0.001$; from 10^5 random permutations of individuals between samples). Taking a closer look at the regeneration sample, we also observed a decreasing trend of immigration rates with recruit age: 0.300 (95% CI: 0.117–0.585) for <10 years ($n_S = 46$), 0.189 (95% CI: 0.112–0.290) for 10–20 years ($n_S = 214$) and 0.113 (95% CI: 0.008–0.416) for 20–30 years ($n_S = 30$), although these estimates were not significantly different from each other ($P > 0.1$). Two main hypotheses with contrasting management implications can be proposed to explain these trends: (i) immigrants sired by pollen donors from the plantations have a post-dispersal selective disadvantage relative to individuals sired by local fathers; and (ii) the plantations' fecundity, and consequently the proportion of non-local male gametes available for immigration into natural stands, has increased during the last years as planted trees began to reach reproductive maturity. If the first ('migration load') hypothesis was true, the natural selection would tend to eliminate individuals sired by non-local fathers in the long term, provided that immigration rates and selection differentials remained unchanged. If the second hypothesis was the main reason for the observed pattern, we could expect even further immigration increments in coming years, which would increase the exposure of the relict population to non-local genes and, ultimately, lead to the eventual displacement of local ones ('gene swamping'). In order to test the migration load hypothesis, it would be necessary to carry out common garden experiments under controlled conditions, measuring the relative performance under varied environments of progenies from local mothers sired by either local or non-local pollen donors. The latter experiment represents a necessary further step, subsequent to the estimation of migration rates and their associated statistical uncertainty, to build a sounder risk assessment protocol for gene flow in conservation biology.

Acknowledgements

We thank the editor and one anonymous reviewer for constructive comments on the paper. This work was supported by Grants 59/2003 and AEG06-054 from the 'Organismo Autónomo de Parques Nacionales' and 'Dirección General para la Biodiversidad', respectively, both belonging to the Spanish 'Ministerio de Medio Ambiente y Medio Rural y Marino'. JJRA and SCGM were supported by Ramón y Cajal research fellowships from the Spanish 'Ministerio de Ciencia e Innovación'.

References

Adams WT, Hipkins VD, Burczyk J, Randall WK (1997). Pollen contamination trends in a maturing Douglas-fir seed orchard. *Can J For Res* **27**: 131–134.

Alía R, Notivol E, Moro J (2001). Genetic variability of Spanish provenances of Scots pine (*Pinus sylvestris* L.): growth traits and survival. *Silva Fenn* **35**: 27–38.

Coles S (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer: London.

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

Farjon A (1998). *World Checklist and Bibliography of Conifers*. Royal Botanic Gardens, Kew.

Faubet P, Gaggiotti OE (2008). A new bayesian method to identify the environmental factors that influence recent migration. *Genetics* **178**: 1491–1504.

Faubet P, Waples RS, Gaggiotti OE (2007). Evaluating the performance of a multilocus Bayesian method for the estimation of migration rates. *Mol Ecol* **16**: 1149–1166.

Hu XS, Ennos RA (1999). Impacts of seed and pollen flow on population genetic structure for plant genomes with three contrasting modes of inheritance. *Genetics* **152**: 441–450.

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

Lopez S, Rousset F, Shaw FH, Shaw RG, Ronce O (2008). Migration load in plants: role of pollen and seed dispersal in heterogeneous landscapes. *J Evolution Biol* **21**: 293–309.

Manel S, Gaggiotti OE, Waples RS (2005). Assignment methods: matching biological questions with appropriate techniques. *Trends Ecol Evol* **20**: 136–142.

Manly BFJ (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 2nd edn. Chapman & Hall: London.

Mercer KL, Wainwright JD (2008). Gene flow from transgenic maize to landraces in Mexico: an analysis. *Agr Ecosyst Environ* **123**: 109–115.

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

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

Pella J, Masuda M (2001). Bayesian methods for analysis of stock mixtures from genetic characters. *Fish bull* **99**: 151–167.

Plomion C, LeProvost G, Pot D, Vendramin G, Gerber S, Decroocq S *et al*. (2001). Pollen contamination in a maritime pine polycross seed orchard and certification of improved seeds using chloroplast microsatellites. *Can J Forest Res* **31**: 1816–1825.

Provan J, Soranzo N, Wilson NJ, McNicol JW, Forrest GI, Cottrell J *et al*. (1998). Gene-pool variation in Caledonian and European Scots pine (*Pinus sylvestris* L.) revealed by chloroplast simple-sequence repeats. *P Roy Soc Lond B Bio* **265**: 1697–1705.

Rannala B, Mountain JL (1997). Detecting immigration by using multilocus genotypes. *P Natl Acad Sci USA* **94**: 9197–9201.

Robledo-Arnuncio JJ, Collada C, Alía R, Gil L (2005). Genetic structure of montane isolates of *Pinus sylvestris* L. in a Mediterranean refugial area. *J Biogeogr* **32**: 595–605.

Sinclair WT, Morman JD, Ennos RA (1999). The postglacial history of Scots pine (*Pinus sylvestris* L.) in western Europe: evidence from mitochondrial DNA variation. *Mol Ecol* **8**: 83–88.

Slavov GT, Howe GT, Gyaourova AV, Birkes DS, Adams WT (2005). Estimating pollen flow using SSR markers and paternity exclusion: accounting for mistyping. *Mol Ecol* **14**: 3109–3121.

Stoehr MU, Newton CR (2002). Evaluation of mating dynamics in a lodgepole pine seed orchard using chloroplast DNA markers. *Can J Forest Res* **32**: 469–476.

Vendramin G, Lelli L, Rossi P, Morgante M (1996). A set of primers for the amplification of 20 chloroplast microsatellites in *Pinaceae*. *Mol Ecol* **5**: 595–598.

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