

# Using Genetic Markers to Directly Estimate Gene Flow and Reproductive Success Parameters in Plants on the Basis of Naturally Regenerated Seedlings

J. Burczyk,<sup>\*,1</sup> W. T. Adams,<sup>†</sup> D. S. Birkes<sup>‡</sup> and I. J. Chybicki<sup>\*</sup>

<sup>\*</sup>*Department of Genetics, Institute of Biology and Environmental Protection, University of Bydgoszcz, 85-064 Bydgoszcz, Poland,*

<sup>†</sup>*Department of Forest Science, Oregon State University, Corvallis, Oregon 97331 and* <sup>‡</sup>*Department of Statistics, Oregon State University, Corvallis, Oregon 97331*

Manuscript received June 13, 2005

Accepted for publication February 15, 2006

## ABSTRACT

Estimating seed and pollen gene flow in plants on the basis of samples of naturally regenerated seedlings can provide much needed information about “realized gene flow,” but seems to be one of the greatest challenges in plant population biology. Traditional parentage methods, because of their inability to discriminate between male and female parentage of seedlings, unless supported by uniparentally inherited markers, are not capable of precisely describing seed and pollen aspects of gene flow realized in seedlings. Here, we describe a maximum-likelihood method for modeling female and male parentage in a local plant population on the basis of genotypic data from naturally established seedlings and when the location and genotypes of all potential parents within the population are known. The method models female and male reproductive success of individuals as a function of factors likely to influence reproductive success (*e.g.*, distance of seed dispersal, distance between mates, and relative fecundity—*i.e.*, female and male selection gradients). The method is designed to account for levels of seed and pollen gene flow into the local population from unsampled adults; therefore, it is well suited to isolated, but also wide-spread natural populations, where extensive seed and pollen dispersal complicates traditional parentage analyses. Computer simulations were performed to evaluate the utility and robustness of the model and estimation procedure and to assess how the exclusion power of genetic markers (isozymes or microsatellites) affects the accuracy of the parameter estimation. In addition, the method was applied to genotypic data collected in Scots pine (isozymes) and oak (microsatellites) populations to obtain preliminary estimates of long-distance seed and pollen gene flow and the patterns of local seed and pollen dispersal in these species.

ONE of the major challenges in plant population biology is to describe the relationship between population genetic structure and reproductive patterns of a species. This objective has often been addressed through parentage analysis using neutral genetic markers (DEVLIN and ELLSTRAND 1990; MEAGHER 1991), and recent availability of highly variable genetic markers, such as microsatellites, has greatly increased the efficiency of parentage assignment (DOW and ASHLEY 1996; STREIFF *et al.* 1999; GONZALEZ-MARTINEZ *et al.* 2002). Parentage inference has now advanced to the point where demonstrating variation in reproductive success among adult plants is no longer sufficient by itself. It is more interesting to identify causes of differential male and female reproductive success. In theory, reproductive patterns are best inferred when parents can be unambiguously assigned through parentage analysis. If, however, genetic discrimination is not complete and/or

errors occur in parentage assignment, estimates of parameters describing reproductive patterns in a population may be seriously biased (ROEDER *et al.* 1989). Under these conditions, the population-wide parameters describing a reproductive system (*e.g.*, selection gradients, see MORGAN and CONNER 2001; BURCZYK *et al.* 2002) are best estimated by applying probability models that account for the genetic composition of offspring in actual progeny arrays and inferring the appropriate model parameters. This method has been used effectively to study male mating success (ADAMS and BIRKES 1991; BURCZYK *et al.* 1996, 2002; SMOUSE *et al.* 1999; MORGAN and CONNER 2001) and female reproductive success (SCHNABEL *et al.* 1998) in plant populations.

Estimation of reproductive patterns in plants is further complicated because gene flow (*i.e.*, immigration of seed and/or pollen gametes) from uncensored parents residing outside the study population often contributes to offspring (ENNOS 1994; HAMRICK and NASON 2000; MORGAN and CONNER 2001). Methods of simultaneously estimating paternity due to pollen gene flow and selection gradients describing differential fertility

<sup>1</sup>*Corresponding author:* University of Bydgoszcz, Institute of Biology and Environmental Protection, Department of Genetics, Chodkiewicza 30, 85-064 Bydgoszcz, Poland. E-mail: burczyk@ukw.edu.pl

of males within populations, by applying mating models to the genetic composition of progeny arrays (seeds) sampled from individual mother plants, have been described and illustrated (ADAMS and BIRKES 1991; BURCZYK *et al.* 1996, 2002; see also SMOUSE *et al.* 1999; MORGAN and CONNER 2001; WRIGHT and MEAGHER 2004). These procedures can be used only to estimate male components of reproduction at the seed stage and when seeds are sampled from individual, known mother plants. Similar methods, however, can be applied to describe gene flow by seeds and female selection gradients (or effective seed dispersal) within populations (ADAMS 1992; DOW and ASHLEY 1996; SCHNABEL *et al.* 1998; GODOY and JORDANO 2001; GONZALEZ-MARTINEZ *et al.* 2002).

It might be expected that realized gene flow by pollen or seeds (MEAGHER and THOMPSON 1987) and selection gradients (female or male fertility) observed in naturally regenerated seedlings may differ substantially from these parameters measured at the seed stage (DYER and SORK 2001). Several factors such as seed dispersal mechanisms, seed predation, and selection during germination and establishment may influence the genetic composition of naturally regenerated seedlings. For example, if immigrating pollen comes from populations adapted to substantially different environmental conditions, the resulting offspring may not be as fit as offspring derived from local matings. Therefore, revealing patterns and determinants of gene dispersal and reproductive success is fundamental to understanding the genetic aspects of natural regeneration in plant populations (MEAGHER and THOMPSON 1987). However, because of the high amount of genetic exclusion power required (MARSHALL *et al.* 1998), there have been only a few attempts to study parentage of naturally established seedlings (MEAGHER and THOMPSON 1987; DOW and ASHLEY 1996; SCHNABEL *et al.* 1998; KONUMA *et al.* 2000; KAMEYAMA *et al.* 2001; GONZALEZ-MARTINEZ *et al.* 2002; ISAGI and KANAZASHI 2002; JONES 2003; SHIMATANI 2004). Nevertheless, these studies were unable to describe fully the reproductive patterns that led to the establishment of the seedling populations.

In this article we present the *seedling neighborhood model* (BURCZYK *et al.* 2004), a novel probability model that makes it possible to describe various reproductive factors (*i.e.*, gene flow and selection gradients) influencing the genetic composition and genealogy of naturally regenerating seedling cohorts. We investigate the statistical properties of the model through simulations, focusing on statistical resolution provided by sets of genetic loci of different exclusion probabilities. Finally, we apply the model to preliminary data sets from Scots pine (*Pinus sylvestris* L.) and oak (*Quercus* sp.) populations. We also discuss potential applications of the model for investigating reproductive patterns in plant populations.

## METHODS

**Seedling neighborhood model:** This probability mating model is fashioned after our earlier neighborhood models developed to describe pollen gene flow and mating patterns within populations (ADAMS and BIRKES 1989, 1991; BURCZYK *et al.* 2002). It requires the mapped locations of all sampled seedlings and all potential reproductive adult males and females within a local population, the multilocus genotypes of all seedlings and adults, and allele frequencies of the same species in surrounding (background) populations. The model allows for simultaneous estimation of seed and pollen immigration levels, along with female and male reproductive success parameters (*i.e.*, relative fertilities and seed and pollen dispersal). For bisexual plants it also accounts for the proportion of selfing among the fraction of seedlings originating from local mothers. We assume in this article that parentage of seedlings is of primary interest, but the procedure applies equally well to seeds (embryos) sampled on the ground or in seed traps.

The general idea of the model is outlined in Figure 1. The model assumes that a seedling is mothered either (i) by an unknown female located outside an arbitrarily defined circular area around the seedling (*i.e.*, the seedling's neighborhood, hence the name of the model) due to seed immigration (with probability  $m_s$ ) or (ii) by a specific local female growing within the seedling's neighborhood (with probability  $1 - m_s$ ). For each seedling with a local mother it is assumed that the paternal gamete came from one of three sources: (i) self-fertilization (with probability  $s$ ), (ii) migrant pollen from outside of the mother's neighborhood (with probability  $m_p$ ), or (iii) outcrossing with males located within the mother's neighborhood (with probability  $1 - s - m_p$ ), much like earlier neighborhood models for pollen dispersal (ADAMS and BIRKES 1991; BURCZYK *et al.* 2002). Note that the mother's neighborhood is a circular area surrounding the mother that has the same radius as the seedling neighborhood. The probability of observing the  $i$ th seedling having a multilocus diploid genotype  $G_i$ , therefore, is

$$P(G_i) = m_s \cdot P(G_i | B_s) + (1 - m_s) \cdot \sum_j \psi_{ij} \cdot [s \cdot P(G_i | M_{ij}, M_{ij}) + m_p \cdot P(G_i | M_{ij}, B_p) + (1 - s - m_p) \cdot \sum_k \phi_{ijk} \cdot P(G_i | M_{ij}, F_{ijk})], \quad (1)$$

where  $M_{ij}$  and  $F_{ijk}$  are the genotypes of the  $j$ th mother in the  $i$ th seedling neighborhood and of the  $k$ th father in the  $ij$ th mother neighborhood, respectively.  $P(G_i | M_{ij}, M_{ij})$ ,  $P(G_i | M_{ij}, B_p)$ , and  $P(G_i | M_{ij}, F_{ijk})$  are the genetic segregation (or *transition*) probabilities (DEVLIN *et al.* 1988), *i.e.*, the probabilities that the  $i$ th seedling has

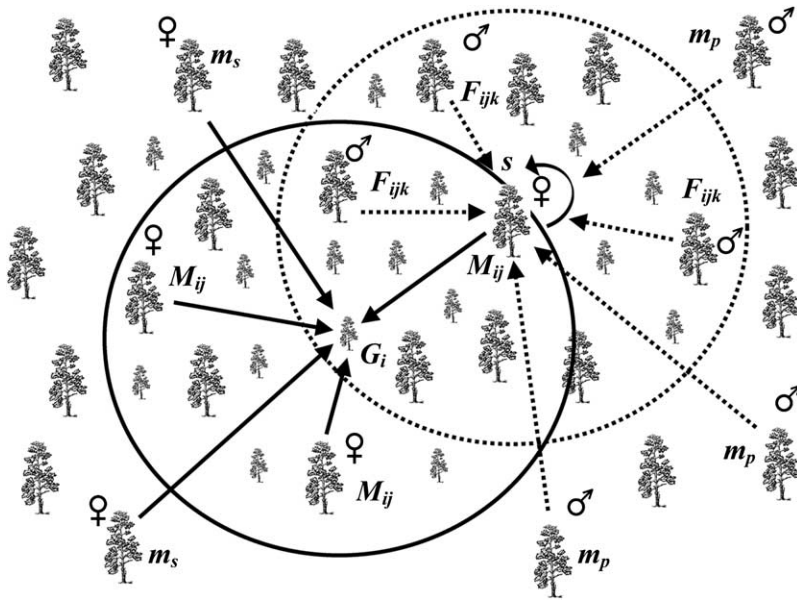


FIGURE 1.—The seedling neighborhood model ascribes female and male parentage of a sampled seedling as follows. The mother of the  $i$ th seedling comes from an unknown female outside the seedling's neighborhood (circle with solid line) (probability  $m_s$ ) or is a specific ( $j$ th) female (*i.e.*,  $M_{ij}$ ) within the neighborhood [probability  $(1 - m_s)\psi_{ij}$ ]. If the mother is the  $j$ th female within the seedling's neighborhood, the male parent comes from one of three sources: by selfing (if bisexual), with probability  $s$ ; from an unknown male outside the mother's neighborhood (circle with dashed line) (*i.e.*, background pollen), with probability  $m_p$ ; or by outcrossing to a specific ( $k$ th) male (*i.e.*,  $F_{ijk}$ ) within the mother's neighborhood, with probability  $(1 - s - m_p)\phi_{ijk}$ . Both seedling and mother neighborhoods have the same radius.

diploid genotype  $G_i$  when a mother plant of genotype  $M_{ij}$  is, respectively, self-pollinated, pollinated by a distant unknown background male, or pollinated by a neighboring plant having genotype  $F_{ijk}$ .  $P(G_i | B_s)$  is the transition probability that a seedling immigrating from mothers located outside of a seedling's neighborhood has genotype  $G_i$ . Parameter  $\psi_{ij}$  is the relative reproductive success of the  $j$ th female in the neighborhood of the  $i$ th seedling, and  $\phi_{ijk}$  is the relative reproductive success of the  $k$ th male within the neighborhood of the  $ij$ th female. In species not capable of self-fertilization  $s = 0$ , and the terms in the model are simplified.

Relative female reproductive success is expressed as

$$\psi_{ij} = \frac{\tau_{ij}}{\sum_h s_i \tau_{ih}}, \tag{2}$$

where  $\tau_{ij}$  is a function of one or more factors influencing the reproductive success of the  $j$ th female in the  $i$ th seedlings neighborhood, and the denominator is the sum over all ( $s_i$ ) potential females in that neighborhood (so that  $\sum \psi_{ij} = 1$ ). Similarly, male reproductive success is

$$\phi_{ijk} = \frac{\pi_{ijk}}{\sum_l z_{ij} \pi_{ijl}}, \tag{3}$$

where  $\pi_{ijk}$  is a function of factors influencing the reproductive success of the  $k$ th male in the neighborhood of the  $j$ th female located within the  $i$ th seedlings neighborhood, and  $z_{ij}$  is the total number of potential fathers in the  $ij$ th female's neighborhood.

Although one may use various types of functions for  $\tau_{ij}$  and  $\pi_{ijk}$  to relate female and male reproductive success to factors influencing reproductive success, here we use an exponential function [ $\tau_{ij} = \exp(\omega_{ij})$ ;  $\pi_{ijk} = \exp(v_{ijk})$ ] because this assures positivity of the reproductive success parameters  $\psi_{ij}$  and  $\phi_{ijk}$ , which can be regarded as

relative proportions. Positivity is maintained for any expression for  $\omega_{ij}$  and  $v_{ijk}$  accommodating various factors affecting reproductive success (BURCZYK *et al.* 1996, 2002; BURCZYK and PRAT 1997; BACLES *et al.* 2005). In particular, seed and pollen dispersal can readily be described through exponential distributions. Variation in parental fitness based on relative fecundity surrogates (*e.g.*, number of flowers, plant size) can also be approximated by an exponential distribution (SMOUSE *et al.* 1999; MORGAN and CONNER 2001). For example, if  $\omega_{ij}$  includes two factors, such as the distance of a seedling from potential mothers within its neighborhood and relative size (*e.g.*, height) of the mothers, then we can let  $\omega_{ij} = \gamma_1 d_{ij} + \gamma_2 f_{ij}$ , where  $d_{ij}$  is the distance between a seedling and a potential mother and  $f_{ij}$  is the mother's relative size. If similar factors are considered for male reproductive success then let  $v_{ijk} = \beta_1 d_{ijk} + \beta_2 f_{ijk}$ , where  $d_{ijk}$  and  $f_{ijk}$  are the distance between male and female and the male's size, respectively. The parameters  $\gamma_1$ ,  $\gamma_2$  and  $\beta_1$ ,  $\beta_2$ , in these cases, describe the strength and direction of the effects of their respective factors. These parameters are often referred to as selection and ecological gradients and they represent the slope of regression of individual fertilities on trait values (see MORGAN and CONNER 2001; BURCZYK *et al.* 2002). The linear functions  $\omega_{ij}$  and  $v_{ijk}$  can be further extended to include quadratic terms, making it possible to assess the effects of stabilizing or disruptive selection acting on particular traits influencing reproductive success (MORGAN and CONNER 2001; WRIGHT and MEAGHER 2004). The unique feature of the seedling neighborhood model is that it can be applied to simultaneously estimating selection gradients of a given phenotypic trait in both male and female parents, making it possible to evaluate the importance of the trait in both male and female reproductive success.

To estimate the model's parameters the model is "fitted" to observed multilocus genotypic data of seedlings and adults using numerical procedures based on maximum-likelihood methods (see ADAMS 1992; BURCZYK *et al.* 1996, 2002 for explanations). The likelihood function for a sample of  $n$  independently selected offspring individuals is

$$L(m_s, m_p, s, \boldsymbol{\gamma}, \boldsymbol{\beta}) = \prod_{i=1}^n P(G_i), \quad (4)$$

where  $\boldsymbol{\gamma}$  and  $\boldsymbol{\beta}$  are the vectors of parameters related to female and male reproductive success, respectively. Stated in nontechnical terms, the following can be estimated using these procedures given sufficient data: (1) the proportion of seedlings established from seed by distant ( $m_s$ ) vs. nearby ( $1 - m_s$ ) females; (2) the degree to which various factors such as distance and relative fecundity influence relative reproductive success of females within a local population; (3) the proportion of a mother's offspring within a local population (at the seedling stage) due to self-fertilization ( $s$ ), pollination by distant males ( $m_p$ ), or pollination by nearby males ( $1 - m_p$ ); and (4) the degree to which various factors such as distance to females and pollen fecundity influence reproductive success of males within local populations.

**Simulations:** We wished to explore the efficiency of the model and estimation procedure through computer simulation, which is frequently used to investigate statistical properties of mating models (MARSHALL *et al.* 1998; MORGAN 1998; GERBER *et al.* 2000; MORGAN and CONNER 2001). Each simulation was initiated by generating a bisexual parental population consisting of 200 adults randomly distributed across a square area of  $150 \times 150$  units. Also, the cohort of 500 seedlings was simulated and randomly distributed in the center of the plot across a square area of  $30 \times 30$  units. In this way, drawing a neighborhood of 30 units in radius around seedlings and then the same size neighborhood around potential females resulted in complete neighborhoods located entirely within the generated plot. The average number of adults within a neighborhood of 30 units in radius was  $\sim 25$  individuals. Distances between seedlings and potential mothers ( $d_{ij}$ ), as well as distances among adults ( $d_{ijk}$ , male-female pairs), were calculated and used as factors influencing female and male reproductive successes ( $\boldsymbol{\gamma}$  and  $\boldsymbol{\beta}$  parameters, respectively).

Each parent was assigned a multilocus genotype by randomly sampling alleles from a specified frequency distribution (see below). Alleles were sampled independently within and among loci, so the parental population was in Hardy-Weinberg equilibrium and with linkage equilibrium among loci. The parental generation and the specified frequency distribution were used to generate seedling genotypes. First, the proportion  $m_s$  of seedlings had their diploid genotypes generated on the basis of the background allelic frequency distribu-

tion (same as for adults), which was assumed to represent immigrant seeds from a background source. For each of the remaining seedlings ( $1 - m_s$ ) a female parent was chosen within the seedling neighborhood using a random number generator, with the probability of choosing the  $j$ th female given by  $\psi_{ij}$  in Equation 2, and a multilocus egg gamete was determined by randomly sampling alleles from the female parent's genotype. For each seedling whose female parent was within the neighborhood, the source of pollen gamete was determined randomly by the following probabilities:  $s$  (self pollen),  $m_p$  (pollen from the background source), and  $1 - s - m_p$  (pollen from a male within the female parent's neighborhood, with the probability of choosing the  $k$ th male given by  $\phi_{ijk}$  in Equation 3). The pollen haplotype from the specific source was then assigned by randomly sampling alleles from that source, and the genotypes of egg and pollen gametes were combined to form the seedling genotype. The complete set of adult and seedling genotypes was then subjected to the estimation procedure described above.

The potentially large number of parameters and their combinations prevent a comprehensive investigation of statistical properties. For simulations, we used two sets of marker loci varying in their exclusion power. The first set (marker set I) included six loci, each with three alleles at frequencies 0.7, 0.2, and 0.1. This set, with exclusion probability (EP) = 0.8034 (CHAKRABORTY *et al.* 1988), might be considered typical for isozymes. The second set (marker set II) included six loci with 10 alleles, each with equal frequency (0.1) with EP = 0.9999, resembling a battery of microsatellites.

First we explored the properties of the model and estimation procedure in the simple case where only levels of seed and pollen immigration ( $m_s$  and  $m_p$ ) were of interest, and there are no selfing ( $s = 0$ ) or differences in reproductive success (female or male) among parents in the population (*i.e.*, parameters related to female and male reproductive success  $\boldsymbol{\gamma} = \boldsymbol{\beta} = 0$ ). We simulated seedling samples on the basis of true parameter values of seed ( $m_s$ ) and pollen immigration ( $m_p$ ) ranging from 0.2 to 0.8, in various combinations (see Table 1). Second, seedling cohorts were simulated assuming that reproductive success of females and males is a function of seed and pollen dispersal within neighborhoods following an exponential distribution (Table 2). Here, the negative values of  $\boldsymbol{\gamma}$  and  $\boldsymbol{\beta}$  were used to represent the decreasing reproductive success with increasing the dispersal distance. The combination of parameters  $\boldsymbol{\gamma} = -0.20$ ,  $\boldsymbol{\beta} = -0.10$ ,  $s = 0.2$  was used to represent a case of severely localized seed and pollen dispersal. Here, mean effective numbers of seed and outcross pollen parents (*i.e.*, selfing excluded) within seedling and mother tree neighborhoods were on average  $N_{es} = 7.08$  and  $N_{ep} = 15.47$ , respectively (see CROW and KIMURA 1970; BURCZYK *et al.* 2002), as compared to the mean census number of  $\sim 25$  individuals. The

TABLE 1

Bias and standard deviations (over 500 replicates, in parentheses) of maximum-likelihood estimates of two immigration parameters ( $m_s$  and  $m_p$ ) in the seedling neighborhood model obtained from computer-simulated data sets for various combinations of  $m_s$  and  $m_p$  and two levels of genetic precision (*i.e.*, where the exclusion probability was either 0.80 or 0.99)

Parameter values		Isozymes EP = 0.80			Microsatellites EP = 0.99		
$m_s$	$m_p$	$\hat{m}_s$	$\hat{m}_p$	$f^a$	$\hat{m}_s$	$\hat{m}_p$	$f^a$
0.2	0.2	-0.0031 (0.0681)	-0.0017 (0.1240)	17	-0.0002 (0.0183)	-0.0006 (0.0203)	0
0.5	0.5	0.0018 (0.0842)	-0.0282 (0.2036) <sup>b</sup>	24	-0.0002 (0.0232)	0.0027 (0.0322)	0
0.8	0.8	-0.0030 (0.0850)	-0.1439 (0.4553) <sup>b</sup>	387	-0.0003 (0.0195)	-0.0012 (0.0404)	0
0.2	0.8	0.0016 (0.0777)	-0.0089 (0.1062)	17	0.0002 (0.0182)	0.0004 (0.0200)	0
0.3	0.7	0.0024 (0.0804)	-0.0167 (0.1292) <sup>b</sup>	11	-0.0001 (0.0211)	-0.0003 (0.0246)	0
0.4	0.6	0.0054 (0.0833)	-0.0069 (0.1609)	15	0.0002 (0.0227)	0.0010 (0.0286)	0
0.6	0.4	0.0021 (0.0847)	-0.0062 (0.2625)	109	0.0000 (0.0229)	-0.0019 (0.0361)	0
0.7	0.3	-0.0079 (0.0839) <sup>b</sup>	0.0081 (0.3503)	290	-0.0002 (0.0217)	0.0007 (0.0410)	0
0.8	0.2	-0.0246 (0.0848) <sup>b</sup>	0.0916 (0.5050) <sup>b</sup>	615	-0.0001 (0.0193)	-0.0019 (0.0492)	1
SE range of simulated bias		0.0030–0.0038	0.0047–0.0226		0.0008–0.0010	0.0009–0.0022	

$m_s$  is the probability that a seedling is the result of seed dispersed from a distant female (outside the neighborhood of the seedling).  $m_p$  is the probability that a seedling is the result of pollination by a distant male (outside the neighborhood of the female parent). In this case, it was assumed that  $s = \gamma = \beta = 0$ . See text for details.

<sup>a</sup> Number of times the estimation procedure failed to converge to biologically reasonable estimates of reproductive parameters is shown. Simulations of data sets were continued until 500 gave reasonable estimates.

<sup>b</sup> Bias significant at  $P < 0.05$ .

parameter combination  $\gamma = -0.10$ ,  $\beta = -0.05$ ,  $s = 0.1$  simulated less restricted seed and pollen dispersal ( $N_{es} = 15.40$ ,  $N_{ep} = 22.03$ ). The seed and pollen immigration levels were set at three categories: (i)  $m_s = m_p = 0$ , fully isolated population with no seed and pollen immigration (*i.e.*, all parents are within neighborhoods); (ii)  $m_s = 0.2$  and  $m_p = 0.5$ , moderate seed and pollen immigration; and (iii)  $m_s = 0.5$  and  $m_p = 0.8$ , extensive seed and pollen immigration. The range of parameter values used for simulations was chosen on the basis of the preliminary analyses of the example data sets (see below).

For each parameter combination 500 simulations were performed. The parameter estimates were obtained numerically on the basis of maximum-likelihood procedures using the Newton–Raphson method (RAO 1973; KENNEDY and GENTLE 1980). Means and variances of the estimates over the replicated data sets were then compared to the true parameter values.

**Example data:** In addition to the simulated data sets, we applied the seedling neighborhood model to actual data obtained from two forest stands: a Scots pine (*P. sylvestris* L.) stand and a mixed-oak [*Quercus robur* L./*Q. petraea* (Matt.) Liebl.] stand. Both are certified seed collection stands located in Poland (Scots pine, Forest District Woziwoda; mixed oaks, Forest District Jamy). In the Scots pine stand, 525 seedlings (5–15 years old) and 313 adults (~160 years old) were genotyped on the basis of eight allozyme loci (estimated EP = 0.72). In the oak stand, 320 seedlings (1–3 years old) and 450 adults (120 years old) were genotyped on the basis of three

microsatellite loci (EP = 0.95). The locations of all seedlings and adults were mapped allowing for a detailed analysis of the effect of seed and pollen dispersal on corresponding male and female reproductive success within each stand. The neighborhood radius around seedlings and putative mothers was set to 40 m in both stands, which included on average 83.8 trees (166/ha) in the Scots pine and 73.9 trees (146.8/ha) in the mixed-oak stands. Such numbers of adults within neighborhoods seemed sufficient for precise estimation of seed and pollen dispersal patterns within neighborhoods (BURCZYK *et al.* 1996). Also, with a 40-m radius, the seedling and subsequent mother neighborhoods were entirely included within the sampled plots of both stands. Reproductive parameters ( $m_s$ ,  $m_p$ ,  $\gamma$ ,  $\beta$ , and  $s$ ) were estimated using the computer program SNM v.1.0, written for this purpose (available from I. J. CHYBICKI upon request), which employs optimized procedures (Powell’s and Newton–Raphson methods) to estimate parameters. The standard deviations of the parameters were derived from the Hessian (variance–covariance) matrix, which is an inherent part of the estimation procedure employed in the SNM program. Additionally, the significance of the parameters  $\gamma$ ,  $\beta$ ,  $s$  was assessed using likelihood-ratio tests (MANLY 1992; MORGAN and CONNER 2001).

RESULTS AND DISCUSSION

**Simulations:** Simulations indicated that the seedling model gives reasonably robust estimates of gene flow

TABLE 2

Bias and standard deviations (over 500 replicates, in parentheses) of maximum-likelihood estimates of neighborhood model parameters obtained from computer-simulated data sets of various combinations of model parameters and two levels of genetic precision (*i.e.*, where the exclusion probability was either 0.80 or 0.99)

Parameter values			Bias (and SD) of parameter estimates						$f^a$	
$m_s$	$m_p$	$s$	$\gamma$	$\beta$	$\hat{m}_s$	$\hat{m}_p$	$\hat{s}$	$\hat{\gamma}$	$\hat{\beta}$	$f^a$
Isozymes EP = 0.80										
0.0	0.0	0.1	-0.1	-0.05	—	—	0.0011 (0.0205)	0.0001 (0.0105)	-0.0001 (0.0138)	0
		0.2	-0.2	-0.10	—	—	-0.0012 (0.0240)	-0.0009 (0.0118)	-0.0002 (0.0127)	0
0.2	0.5	0.1	-0.1	-0.05	-0.0024 (0.0674)	-0.0087 (0.1082)	-0.0003 (0.0257)	-0.0007 (0.0146)	0.0027 (0.0387)	6
		0.2	-0.2	-0.10	-0.0039 (0.0525)	0.0050 (0.0834)	0.0026 (0.0308)	-0.0010 (0.0174)	-0.0299 (0.1056) <sup>b</sup>	8
0.5	0.8	0.1	-0.1	-0.05	0.0058 (0.0809)	-0.0296 (0.1537) <sup>b</sup>	-0.0013 (0.0377)	-0.0017 (0.0229)	— <sup>c</sup>	90
		0.2	-0.2	-0.10	-0.0018 (0.0664)	-0.0081 (0.1199)	-0.0025 (0.0337)	-0.0134 (0.0322) <sup>b</sup>	— <sup>c</sup>	46
Microsatellites EP = 0.99										
0.0	0.0	0.1	-0.1	-0.05	—	—	-0.0008 (0.0133)	0.0000 (0.0070)	0.0005 (0.0065)	0
		0.2	-0.2	-0.10	—	—	-0.0007 (0.0178)	-0.0004 (0.0085)	-0.0004 (0.0071)	0
0.2	0.5	0.1	-0.1	-0.05	-0.0001 (0.0183)	-0.0003 (0.0252)	0.0002 (0.0150)	-0.0002 (0.0074)	0.0007 (0.0110)	0
		0.2	-0.2	-0.10	0.0003 (0.0183)	-0.0007 (0.0251)	-0.0002 (0.0200)	-0.0007 (0.0093)	-0.0003 (0.0129)	0
0.5	0.8	0.1	-0.1	-0.05	0.0003 (0.0232)	-0.0010 (0.0255)	0.0007 (0.0190)	-0.0002 (0.0092)	-0.0021 (0.0286)	2
		0.2	-0.2	-0.10	-0.0005 (0.0231)	-0.0004 (0.0255)	0.0014 (0.0190)	0.0008 (0.0120)	-0.0023 (0.0297)	1

$m_s$  is the proportion of seed immigrants,  $m_p$  is the proportion of pollen immigrants,  $s$  is the proportion of self-fertilization,  $\gamma$  and  $\beta$  are parameters of female and male reproductive success, respectively (see text for details).

<sup>a</sup> Number of times the estimation procedure failed to converge to biologically reasonable estimates of reproductive parameters is shown. Simulations of data sets were continued until 500 gave reasonable estimates.

<sup>b</sup> Bias significant at  $P < 0.05$ .

<sup>c</sup> The model could not converge with parameter  $\beta$  included in the model.

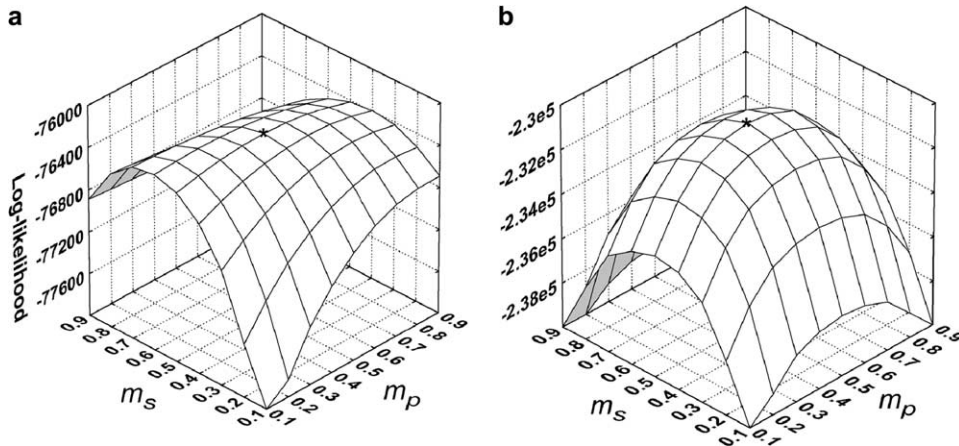


FIGURE 2.—The approximated log-likelihood surface obtained for a simulated data set based on a large seedling sample ( $n = 10,000$ ) when the true parameter values for seed ( $m_s$ ) and pollen ( $m_p$ ) immigration are  $m_s = m_p = 0.5$ : (a) low exclusion power genetic markers ( $EP = 0.80$ ); (b) high exclusion power genetic markers ( $EP = 0.99$ ).

and reproductive parameters when genetic markers with at least moderately high exclusion probabilities ( $EP \geq 0.8$ ) are available (Tables 1 and 2). When individuals within local populations did not differ in female or male reproductive success ( $s = \gamma = \beta = 0$ ), estimates of  $m_s$  and  $m_p$  were nearly unbiased; however, the marker set I (isozyme-like) showed high variances of the estimates (Table 1). Notably, while variances of  $\hat{m}_s$  were stable across a range of true  $m_s$  values, the variances of  $\hat{m}_p$  tended to increase with increasing  $m_s$ . This is because higher  $m_s$  reduces the number of nonimmigrant seedlings contributing to estimating pollen immigration. The standard deviations of  $\hat{m}_p$  for marker set I appeared to be unacceptably high in most cases, given the sample size used in the simulations ( $n = 500$ ). However, it is expected that increased sample size will, at least, partly compensate for a lower exclusion probability (MORGAN 1998; MORGAN and CONNER 2001). For marker set II (microsatellite type), the parameter estimates were unbiased and the variances of both  $\hat{m}_s$  and  $\hat{m}_p$  were much lower, although still being larger for  $\hat{m}_p$ .

With some exceptions (marker set I), the seedling model and estimation procedure provided efficient estimates (low bias and variance) in the cases where the simulations assumed reproductive success of individuals within local populations decreased exponentially with distance from the seedling location (females) or mate (males) (Table 2). When we assumed that all seed and pollen parents are local ( $m_s = m_p = 0$ ), the variances of  $\hat{\gamma}$  and  $\hat{\beta}$  were low and of equivalent magnitude, within a given marker set. However, the variances of both estimates tended to increase with increasing  $m_s$  and  $m_p$ . In all cases, mean parameter estimates of both seed and pollen immigration were very close to their true values (*i.e.*, low bias). The means of standard deviations of parameters based on the Hessian matrix for each simulated data set were nearly identical to standard deviations derived from replicate simulations.

The estimation procedure often failed to converge to reasonable parameter estimates for the simplified data sets with the lower EP, probably due to insufficient

genetic resolution. The Newton–Raphson algorithm used in our iterative estimation procedure requires careful choice of initial parameter values for convergence, especially for data sets with less genetic information (THISTED 1988; MORGAN and CONNER 2001). In our simulations, the parameter values were used as the initial starting points in the iterations. However, the actual parameter values of individual simulated data sets could differ from the expected  $m_s$  and  $m_p$ , and if such difference was considerable this led to the lack of convergence, more often for marker set I. We investigated the probability surface (expressed as the log-likelihood) around the true parameter values in the special case where  $m_s = m_p = 0.5$  and  $s = \gamma = \beta = 0$ . To obtain smooth graphs, we used a large seedling sample ( $n = 10,000$ ) (Figure 2). In the case of marker set II (microsatellites), the log-likelihood has a strong peak at the true values of  $m_s$  and  $m_p$  (0.50, 0.50). Here, the starting points, if shifted away from the true parameter values, should still converge closely to expected parameter values. However, for marker set I (isozymes), although the peak of the log-likelihood surface corresponds to the expected parameter values, there is a ridge through the maximum along which the gradient is close to zero and the algorithm could stop at any point on this ridge, depending on the convergence criterion. The starting points, if shifted away from the true values, could easily converge to a point on the log-likelihood surface that is not the peak (*i.e.*, where  $\hat{m}_s$  and/or  $\hat{m}_p$  are not close to their true values), especially for small data sets with less genetic information. Note that the slope of the log-likelihood surface is steeper along the  $m_s$  axis than along  $m_p$  (Figure 2). The distribution of the log-likelihood around the true parameter values along the  $m_s$  and  $m_p$  axes supports our earlier observations, that the variances of  $\hat{m}_s$  are smaller than those of  $\hat{m}_p$  and that these variances are smaller for the microsatellite- than for the isozyme-type genetic markers.

**Empirical seed and pollen gene flow estimates:** The level of seed immigration in the Scots pine population appears considerable; it is estimated that nearly 45% of

TABLE 3

Estimates of reproductive parameters (SE in parentheses) for a Scots pine (*Pinus sylvestris*) stand and a mixed-oak (*Quercus* sp.) stand obtained by applying the seedling neighborhood model and maximum-likelihood estimation procedure to samples of naturally established seedlings

Reproductive parameter	<i>Pinus sylvestris</i>	<i>Quercus</i> sp.
Seed immigration ( $\hat{m}_s$ )	0.4464 (0.1543)	0.0630 (0.0251)
Pollen immigration ( $\hat{m}_p$ )	0.9225 (0.3936)	0.6236 (0.0483)
Seed dispersal effect ( $\hat{\gamma}$ )	-0.0376 (0.0259)	-0.2027 (0.0124)
Pollen dispersal effect ( $\hat{\beta}$ )	— <sup>a</sup>	-0.0908 (0.0177)
Selfing ( $\hat{s}$ )	0.0571 (0.0454)	0.0120 (0.0126)
Seedling sample size	524	312
Genetic marker (exclusion probability)	Isozymes (EP = 0.72)	Microsatellites (EP = 0.92)

<sup>a</sup>Parameter not included in the model.

seedlings resulted from seed flow over distances >40 m (Table 3). In addition, it is estimated that 92% of the seedlings with local mothers resulted from pollination by distant (>40 m) males. These results suggest that only a small percentage ( $0.55 \cdot (1 - 0.92) \cdot 100 = 4.4\%$ ) of the seedlings in this population have both parents that are local. Although  $\hat{m}_s$  and  $\hat{m}_p$  are quite different from zero, their estimated variances are considerable, as are the variances of  $\hat{\gamma}$  and  $\hat{s}$ . The pollen dispersal parameter,  $\hat{\beta}$ , could not be estimated, probably due to low frequency of local pollination events (4.4%). Despite the large number of seedlings sampled (525), the precision of these data was greatly compressed by the low exclusion power (EP = 0.72) of the isozyme data set. Nevertheless, the high observed levels of pollen and seed gene flow are not surprising for Scots pine. This species is well known for extensive seed and pollen dispersal. Its pollen sedimentation velocity is among the lowest in conifers and the species has the greatest capability to disperse seeds among temperate forest trees (GEBUREK 2005). Independent estimates of seed and pollen gene flow obtained in the same population on the basis of seed samples collected in seed traps and directly from mother trees are quite comparable to the results reported above (A. DZIALUK and J. BURCZYK, unpublished results).

In the case of the oak stand, we were able to estimate all intended reproductive parameters, but the patterns of gene flow were much different from those estimated for Scots pine. Only ~6% of seedlings originated from mother trees located >40 m from seedlings (Table 3). Also, seed dispersal within the local population was very limited ( $\hat{\gamma} = -0.2027$ ), which means that the majority of seedlings originated from nearby mother trees. On the other hand, pollen gene flow was extensive, as ~62% of pollen gametes fertilizing local mother trees came from unsampled males situated outside the mothers' neighborhoods. This suggests that a considerable percentage ( $0.94 \cdot (1 - 0.62) \cdot 100 = 35.7\%$ ) of the seedlings in this population have both parents that are local. In addition,

while pollen dispersal within neighborhoods followed a negative exponential distribution, it was less restrictive ( $\hat{\beta} = -0.0908$ ) than seed dispersal (Figure 3). Variances of parameter estimates were low, indicating high precision of the estimation procedure. This was possible because the applied set of microsatellites had relatively high exclusion power (EP = 0.92). The limited seed but extensive pollen dispersal in oaks is consistent with observations in several earlier studies (DOW and ASHLEY 1996; STREIFF *et al.* 1999).

**Conclusions:** The particular problem in parentage assignment for naturally regenerated seedlings is that it is possible to assign only a fraction of the seedlings to unique potential parental pairs (MEAGHER and THOMPSON 1987; CHAKRABORTY *et al.* 1988). This difficulty is due to insufficient exclusion power of genetic markers and/or parentage with unknown (not genotyped) female and/or male parents (seed and pollen immigration). The two-parent problem, *i.e.*, the problem of determining

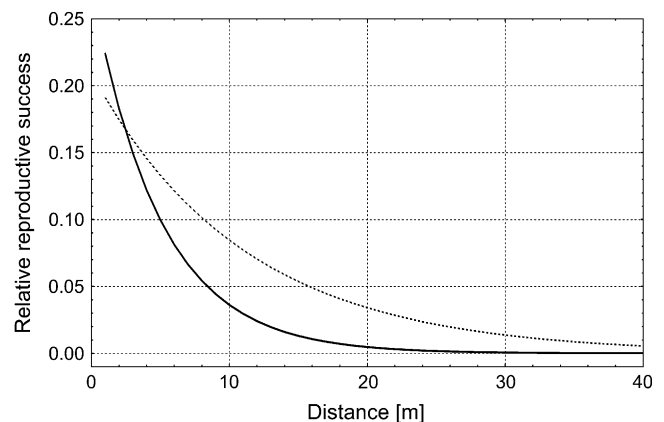


FIGURE 3.—Estimated relationships between relative reproductive success of females and their distance to seedlings ( $\hat{\gamma}$ , solid line) and between relative reproductive success of males and their distance to females ( $\hat{\beta}$ , dotted line) within a local stand (*i.e.*, within neighborhoods) of *Quercus petraea*/ *Q. robur* in Poland. See text for details.



which of the two identified parents is mother and which is father, also complicates parentage analyses, especially in monoecious species (MEAGHER and THOMPSON 1986). The seedling neighborhood model described in this article attempts to overcome these difficulties. The seedling neighborhood model does not assign parentage for offspring but rather models the genealogy of seedlings by reconstructing the two-stage movement of genes through pollen and seeds.

In our description of the model, we defined  $1 - m_s$  as the proportion of seedlings whose mothers are within the seedling's neighborhood. In practice, however, one cannot distinguish between the genotypes of male and female parents of a seedling (two-parent problem; MEAGHER and THOMPSON 1986). It may happen that in real situations the actual mother of a given seedling could be located outside its neighborhood while the father is located within the immediate neighborhood of the seedling. This may lead to underestimation of seed movement and at the same time to overestimation of pollen movement. Nevertheless, if for a given species seed dispersal is more restricted than pollen dispersal, which is typical in plants (ENNOS 1994; HAMRICK and NASON 2000), the populationwide  $m_s$  estimates should approximate the actual proportion of seed immigration.

In this article, we assumed that background allele frequencies are the same for seed and pollen immigration. However, despite little spatial variation in the genetic structure in several natural populations, different modes of seed and pollen dispersal may cause allele frequencies to differ for immigrating seed and pollen pools. Nevertheless, sensitivity of the model to biased background allelic frequencies used in the estimation procedure (BURCZYK and CHYBICKI 2004) decreases with increased exclusion probabilities of the markers. We also assumed that there are no mutations or genotyping errors. However, it might be expected that especially genotyping errors may considerably affect the estimates of seed or pollen immigration (MARSHALL *et al.* 1998; BURCZYK *et al.* 2004; SLAVOV *et al.* 2005). We will address this issue in a future work.

In our simulations and data examples we did not estimate selection gradients (*sensu* MORGAN and CONNER 2001). However, the seedling model can easily accommodate additional parameters that would relate individual reproductive success to individual features, such as plant size, flower fecundity, phenology, or floral characteristics (BURCZYK and PRAT 1997; SMOUSE *et al.* 1999; WRIGHT and MEAGHER 2004). In addition, application of the seedling model makes it possible to assess how these traits affect female *vs.* male reproductive success. Another unique feature of the model is that it is possible to estimate rates of selfing at the seedling stage, which may provide new insights into mechanisms determining inbreeding levels in naturally regenerated populations. We believe that applications of this approach will lead to enhanced recognition of the interactions among

genetics, ecology, and demography at the population level. Such information is of primary interest in genetic conservation programs, where natural regeneration is the major mode of reproduction.

We thank Artur Dzialuk and Magdalena Trojankiewicz for assistance in collecting genotypic data of Scots pine and oak populations. This work was partly supported by research grants from Polish Committee for Scientific Research: 5P06H 042 15 and 3P06L 034 23.

#### LITERATURE CITED

- ADAMS, W. T., 1992 Gene dispersal within forest tree populations. *New For.* **6**: 217–240.
- ADAMS, W. T., and D. S. BIRKES, 1989 Mating patterns in seed orchards, pp. 75–86 in *Proceedings of the 20th Southern Forest Tree Improvement Conference*. Southern Forest Tree Improvement Committee, Charleston, SC.
- ADAMS, W. T., and D. S. BIRKES, 1991 Estimating mating patterns in forest tree populations, pp. 157–172 in *Biochemical Markers in the Population Genetics of Forest Trees*, edited by S. FINESCHI, M. E. MALVOLTI, F. CANNATA and H. H. HATTEMER. SPB Academic Publishing, The Hague, The Netherlands.
- BACLES, C. F. E., J. BURCZYK, A. J. LOWE and R. A. ENNOS, 2005 Historical and contemporary mating patterns in remnant populations of the forest tree *Fraxinus excelsior* L. *Evolution* **59**: 979–990.
- BURCZYK, J., and I. J. CHYBICKI, 2004 Cautions on direct gene flow estimation in plant populations. *Evolution* **58**: 956–963.
- BURCZYK, J., and D. PRAT, 1997 Male reproductive success in *Pseudotsuga menziesii* (Mirb.) Franco: the effect of spatial structure and flowering characteristics. *Heredity* **79**: 638–647.
- BURCZYK, J., W. T. ADAMS and J. Y. SHIMIZU, 1996 Mating patterns and pollen dispersal in a natural knobcone pine (*Pinus attenuata* Lemmon) stand. *Heredity* **77**: 251–260.
- BURCZYK, J., W. T. ADAMS, G. F. MORAN and A. R. GRIFFIN, 2002 Complex patterns of mating revealed in a *Eucalyptus regnans* seed orchard using allozyme markers and the neighborhood model. *Mol. Ecol.* **11**: 2379–2391.
- BURCZYK, J., S. P. DIFAZIO and W. T. ADAMS, 2004 Gene flow in forest trees: How far do genes really travel? *For. Genet.* **11**: 179–192.
- CHAKRABORTY, R., T. R. MEAGHER and P. E. SMOUSE, 1988 Parentage analysis with genetic markers in natural populations. I. The expected proportion of offspring with unambiguous paternity. *Genetics* **118**: 527–536.
- CROW, J. F., and M. KIMURA, 1970 *An Introduction to Population Genetics Theory*. Harper & Row, New York.
- DEVLIN, B., and N. C. ELLSTRAND, 1990 The development and application of a refined method for estimating gene flow from angiosperm paternity analysis. *Evolution* **44**: 248–259.
- DEVLIN, B., K. ROEDER and N. C. ELLSTRAND, 1988 Fractional paternity assignment: theoretical development and comparison to other methods. *Theor. Appl. Genet.* **76**: 369–380.
- DOW, B. D., and M. V. ASHLEY, 1996 Microsatellite analysis of seed dispersal and parentage of saplings in bur oak, *Quercus macrocarpa*. *Mol. Ecol.* **5**: 120–130.
- DYER, R. J., and V. L. SORK, 2001 Pollen pool heterogeneity in short-leaf pine, *Pinus echinata* Mill. *Mol. Ecol.* **10**: 859–866.
- ENNOS, R. A., 1994 Estimating the relative rates of pollen and seed migration among plant populations. *Heredity* **72**: 250–259.
- GEBUREK, TH., 2005 Sexual reproduction in forest trees, pp. 171–198 in *Conservation and Management of Forest Genetic Resources in Europe*, edited by TH. GEBUREK and J. TUROK. Arbora Publishers, Zvolen, Slovakia.
- GERBER, S., S. MARIETTE, R. STREIFF, C. BODENES and A. KREMER, 2000 Comparison of microsatellites and amplified fragment length polymorphism markers for parentage analysis. *Mol. Ecol.* **9**: 1037–1048.
- GODOY, J. A., and P. JORDANO, 2001 Seed dispersal: exact identification of source trees with endocarp DNA microsatellites. *Mol. Ecol.* **10**: 2275–2283.
- GONZALEZ-MARTINEZ, S. C., S. GERBER, M. T. CERVERA, J. M. MARTINEZ-ZAPATER, L. GIL *et al.*, 2002 Seed gene flow and fine-scale structure in a Mediterranean pine (*Pinus pinaster* Ait.) using nuclear microsatellite markers. *Theor. Appl. Genet.* **104**: 1290–1297.

- HAMRICK, J. L., and J. D. NASON, 2000 Gene flow in forest trees, pp. 81–90 in *Forest Conservation Genetics: Principles and Practice*, edited by A. YOUNG, D. BOSHER and T. BOYLE. CSIRO Publishing, Collingwood, Australia.
- ISAGI, Y., and T. KANAZASHI, 2002 Gene flow analysis in *Magnolia obovata* Thunb. using highly variable microsatellite markers, pp. 257–269 in *Diversity and Interaction in a Temperate Forest Community: Ogawa Forest Reserve of Japan*, edited by N. MATSUMOTO. Springer-Verlag, Tokyo.
- JONES, B., 2003 Maximum likelihood inference for seed and pollen dispersal distributions. *J. Agric. Biol. Environ. Stat.* **8**(2): 170–183.
- KAMEYAMA, Y., Y. ISAGI and N. NAKAGOSHI, 2001 Patterns and levels of gene flow in *Rhododendron metternichii* var. *hondoense* revealed by microsatellite analysis. *Mol. Ecol.* **10**: 205–216.
- KENNEDY, JR., W. J., and J. E. GENTLE, 1980 *Statistical Computing*. Marcel Dekker, New York.
- KONUMA, A., Y. TSUMURA, C. T. LEE, S. L. LEE and T. OKUDA, 2000 Estimation of gene flow in the tropical-rainforest tree *Neobalanocarpus heimii* (Dipterocarpaceae), inferred from paternity analysis. *Mol. Ecol.* **9**: 1843–1852.
- MANLY, B. F. J., 1992 *The Design and Analysis of Research Studies*. Cambridge University Press, Cambridge, UK.
- MARSHALL, T. C., J. SLATE, L. E. KRUUK and J. M. PEMBERTON, 1998 Statistical confidence for likelihood-based paternity inference in natural populations. *Mol. Ecol.* **7**: 639–655.
- MEAGHER, T. R., 1991 Analysis of paternity within a natural population of *Chamaelirium luteum*. II. Patterns of male reproductive success. *Am. Nat.* **137**: 738–752.
- MEAGHER, T. R., and E. THOMPSON, 1986 The relationship between single parent and parent pair genetic likelihoods in genealogy reconstruction. *Theor. Popul. Biol.* **29**: 87–106.
- MEAGHER, T. R., and E. THOMPSON, 1987 Analysis of parentage for naturally established seedlings of *Chamaelirium luteum* (Liliaceae). *Ecology* **68**: 803–812.
- MORGAN, M. T., 1998 Properties of maximum likelihood male fertility estimation in plant populations. *Genetics* **149**: 1099–1103.
- MORGAN, M. T., and J. K. CONNER, 2001 Using genetic markers to directly estimate male selection gradients. *Evolution* **55**: 272–281.
- RAO, C. R., 1973 *Linear Statistical Inference and Its Applications*, Ed. 2. Wiley & Sons, New York.
- ROEDER, K., B. DEVLIN and B. G. LINDSAY, 1989 Application of maximum likelihood methods to population genetic data for the estimation of individual fertilities. *Biometrics* **45**: 363–379.
- SCHNABEL, A., J. D. NASON and J. L. HAMRICK, 1998 Understanding the population genetic structure of *Gleditsia triacanthos* L.: seed dispersal and variation in female reproductive success. *Mol. Ecol.* **7**: 819–832.
- SHIMATANI, K., 2004 Spatial molecular ecological models for genotyped adults and offspring. *Ecol. Model.* **174**: 401–410.
- SLAVOV, G. T., G. T. HOWE, D. S. BIRKES, A. V. GYAUROVA and W. T. ADAMS, 2005 Estimating pollen flow using SSR markers and paternity exclusion: accounting for mistyping. *Mol. Ecol.* **14**: 3109–3121.
- SMOUSE, P. E., T. MEAGHER and C. KOBAC, 1999 Parentage analysis in *Chamaelirium luteum* (L.) Gray (Liliaceae): Why do some males have higher reproductive contributions? *J. Evol. Biol.* **12**: 1069–1077.
- STREIFF, R., A. DUCOUSSO, C. LEXER, H. STEINKELLNER, J. GLOESSL *et al.*, 1999 Pollen dispersal inferred from paternity analysis in a mixed oak stand of *Quercus robur* L. and *Q. petraea* (Matt.). *Liebl. Mol. Ecol.* **8**: 831–841.
- THISTED, R. A., 1988 *Elements of Statistical Computing*. Chapman & Hall, New York.
- WRIGHT, J. W., and T. R. MEAGHER, 2004 Selection on floral characters in natural Spanish populations of *Silene latifolia*. *J. Evol. Biol.* **17**: 382–395.

Communicating editor: M. K. UYENOYAMA