MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2012) doi: 10.1111/1755-0998.12005

PSA: software for parental structure analysis of seed or seedling patches

J. J. ROBLEDO-ARNUNCIO,* D. GRIVET,* P. E. SMOUSE† and V. L. SORK‡

*Department of Forest Ecology and Genetics, INIA – Forest Research Centre (CIFOR), Ctra. de la Coruña km 7.5, 28040, Madrid, Spain, †Department of Ecology, Evolution and Natural Resources, Rutgers University, New Brunswick, NJ 08901, USA, ‡Department of Ecology and Evolutionary Biology and Institute of the Environment and Sustainability, University of California, Los Angeles, CA 90095-7239, USA

Abstract

Parental structure analysis (PSA) is a computer program to analyse separate contributions of paternal and maternal parents to postdispersal plant offspring. The program provides joint estimates of maternal, paternal and cross-parental correlations within and among a set of predefined groups of seeds or seedlings, as well as derivative estimates of effective parental numbers. PSA utilizes data sets that distinguish between maternal and paternal contributions to the genotype of each offspring in the sample, but does not require parental samples *per se*. The approach requires assay of codominant diploid markers from both seed coat (maternally inherited) and seedling/embryo (biparentally inherited) tissues for each offspring. A simulation analysis of PSA's performance shows that it provides fairly accurate parental correlation estimates from affordable sampling effort. PSA should be of interest to plant biologists studying the interplay between dispersal, demography and genetics, as well as plant–animal interactions.

Keywords: gene flow, genetic structure, kinship, parentage, pericarp, seed and pollen dispersal

Received 13 June 2012; revision received 1 August 2012; accepted 2 August 2012

Introduction

The genetic structure of plant populations is mediated by the sequential processes of pollen movement, seed transport and seedling establishment. Even if pollen flow is extensive, restricted seed dispersal may create significant spatial genetic structure, while if the converse is true, less fine-scale genetic structure is expected (Grivet *et al.* 2009). One can characterize the long-term (across multiple generations) joint effect of pollen and seed dispersal on spatial genetic structure, via isolation-by-distance analysis (Hardy & Vekemans 2002; Rousset 2008; Rousset & Leblois 2012). For those who study the ecological consequences of propagule dispersal or plant–animal interactions, or who are interested in the genetic consequences of recent demographic, environmental and landscape changes, the separate contemporary contributions of paternal and maternal parents to offspring cohorts may be of greater interest. Parentage-based methods can yield contemporary pollen and seed dispersal kernels, along with estimates of male and female reproductive success, based on the genotypes of a sample of established seedlings (e.g. Burczyk *et al.* 2006; Goto *et al.* 2006), though they also require an exhaustive collection of candidate parents within the study area, which may impose some practical scale limitations. Grivet *et al.* (2009) have proposed alternative analytical methods to infer how contemporary reproductive and dispersal processes contribute to the details of fine-scale spatial genetic structure. Their approach establishes a formal partition of the total effective number of parents contributing to offspring patches into its paternal and maternal components, which are defined in terms of parental correlations, estimated with genetic kinship coefficients.

Unlike parentage analysis, the approach of Grivet *et al.* (2009) does not require parental samples, and it can be applied to data sets in which it is possible to discriminate the male and female gametes of diploid offspring genotypes. Such gametic phase resolution becomes possible when some form of mixed-tissue assay is available, defined as a combination of both a maternally and a biparentally inherited tissue for every offspring in the sample (Grivet *et al.* 2009; Smouse *et al.* 2012). The biparentally inherited diploid genotype of an offspring can be recovered from seedling leaf tissue or seed embryo tissue, while the maternally inherited genotype can be recovered from either the diploid seed pericarp (Godoy & Jordano 2001) or the haploid seed megagametophyte

Correspondence: Juan J. Robledo-Arnuncio, Fax: +34 91 357 2293; E-mail: jjrobledo@gmail.com

(in conifers; Ziegenhagen *et al.* 2003; Iwaizumi *et al.* 2007). The approach can thus be applied to either dispersed seeds (collected from the ground or from seed traps) or naturally established seedlings. In the case of seedlings, it is required that the maternal seed tissue remains attached to the seedling long enough after germination (Grivet *et al.* 2005). Note that, although not considered here, parental and maternal correlations could alternatively be estimated from biparentally inherited tissue alone if polymorphic markers with contrasting modes of inheritance (e.g. paternal/maternal or paternal/biparental) were available, though in that case the estimation of cross-parental correlations (defined in Table 1) would be difficult.

Mixed-tissue assays should be possible for a broad range of species. For taxa with reserve cotyledons (i.e. nonphotosynthetic cotyledons that remain inside the seed, and are called cryptocolar cotyledons) the seeds will stay attached to the seedlings longer, at or slightly below the soil surface (Baskin & Baskin 2000). Storage organs exist generally in large seeds, many of which are buried underground by animals or dispersed by gravity at ground level. A strong association has been found between large seed size and cryptocotyle (Fenner 1992). Storage cotyledons located inside the seed coat at ground level exist in oaks (e.g. Thomas 2000) and in many species producing nuts (e.g. walnut, horse chestnut, cherry, hazel; see examples in Vander Wall 2001), as well as in numerous tropical taxa (e.g. Anacardiaceae, Arecaceae, Fabaceae, Lauraceae, Meliaceae, Myrtaceae, Sapindaceae; see examples in Ibarra-Manríquez *et al.* 2001). So far, seed dispersal studies using this approach have been conducted in seeds as small as those of *Prunus mahaleb* (Godoy & Jordano 2001) and as large as *Oenocarpus bataua*, a palm nut dispersed by umbrellabirds in Ecuador (Karubian *et al.* 2010).

While there are several software implementations to conduct parentage analysis (e.g. Marshall *et al.* 1998; Gerber *et al.* 2003; Chybicki & Burczyk 2010) or isolation-by-distance inference (Hardy & Vekemans 2002; Rousset 2008; Rousset & Leblois 2012), there is currently no computer program available to conduct the parental structure analysis (PSA) introduced in Grivet *et al.* (2009). Programs exist to estimate kinship coefficients among individual pairs (e.g. Hardy & Vekemans 2002; Kalinowski *et al.* 2006; Wang 2011) or conduct sibship reconstruction within groups of individuals (Jones & Wang 2010; see also a review in Blouin 2003), but none of them take advantage of mixed-tissue assay in resolving the gametic phase of diploid individuals, nor do they provide joint estimates of maternal, paternal and cross-parental correlations, as well as derivative estimates of effective parental numbers,

Table 1 Definition and estimation of parameters, as implemented in the parental structure analysis software. Parental correlations are defined as probabilities of particular events (*P*[*event*]) and estimated as twice the average genetic kinship coefficient (F_{ij}) over particular pairs of haplotypes or genotypes (third column) using the samples indicated in the last column to compute reference allelic frequencies

Statistic	Definition	F_{ij} average	Reference sample
Q_w^p, within-patch paternal correlation	*P*[two randomly drawn offspring from a patch have been sired by the same father]	Within-patch pairs of paternal genotypes	Inferred paternal genotypes
Q_w^m, within-patch maternal correlation	*P*[two randomly drawn offspring from a patch have been dispersed from the same mother]	Within-patch pairs of maternal genotypes	Maternal genotypes
Q_w^{mp}, within-patch cross-parental correlation	*P*[two randomly drawn offspring from a patch show a cross-parental match, that is, the mother of the first is the father of the second, or vice versa]	Within-patch pairs of maternal-paternal gametic phases	Biparentally inherited leaf (or embryo) genotypes
Q_b^p, among-patch paternal correlation	*P*[two randomly drawn offspring from two different patches have been sired by the same father]	Pairs of paternal genotypes, one from each patch	Inferred paternal genotypes
Q_b^m, among-patch maternal correlation	*P*[two randomly drawn offspring from two different patches have been dispersed from the same mother]	Pairs of maternal genotypes, one from each patch	Maternal genotypes
Q_b^{mp}, among-patch cross-parental correlation	*P*[two randomly drawn offspring from two different patches show a cross-parental match, that is, the mother of the first is the father of the second or vice versa]	Pairs of maternal–paternal gametic phases, one from each patch	Biparentally inherited leaf (or embryo) genotypes
N_{ep}, effective number of fathers per offspring patch	$1/Q_w^p$		
N_{em}, effective number of mothers per offspring patch	$1/Q_w^m$		
N_e, effective number of parents per offspring patch	$4/(Q_w^p + Q_w^m + 2Q_w^{mp})$		

within and among plant offspring groups. We here introduce PSA, a computer program that conducts such analysis. PSA is available as a Microsoft Windows executable file at https://sites.google.com/site/jjrobledo2/software. The necessary documentation to use the program is provided with the software; this note is intended to: (i) summarize the program functions briefly; (ii) conduct a numerical analysis of the method's performance; and (iii) provide some sampling recommendations.

Program functions

The program PSA assumes a diploid monoecious plant species and requires a mixed-tissue assay of seedling or seed samples collected from a set of predefined patches, using codominant diploid markers such as nuclear microsatellites (SSRs) or single-nucleotide polymorphisms (SNPs). The current version of the program assumes that the available seed maternal tissue is diploid, but extensions to account for haploid maternal tissue would be straightforward. For every locus, PSA obtains the male gametic contribution to each offspring by subtracting the maternal gametic combination from the offspring genotype. This can be performed unambiguously by comparing the maternally and biparentally inherited diploid genotypes, unless they share the same heterozygous state, in which case paternity is assigned fractionally to each of the two possible alleles according to their posterior likelihood value, given the global paternal allelic frequencies estimated from the unambiguous cases (as in Irwin *et al.* 2003; see also Smouse *et al.* 2012).

Once the gametic phase of the diploid genotype of each offspring has been resolved, PSA proceeds to estimate the correlation of maternity, correlation of paternity and cross-parental correlation within and among the sampled offspring patches (see definitions in Table 1). Parental correlations are estimated as twice the average genetic kinship coefficient (F_{ij}) among particular pairs of seedling (or seed) haplotypes or genotypes, assuming unrelated parents (see Table 1 and Grivet *et al.* 2009). PSA implements Loiselle *et al.*'s (1995) kinship coefficient estimator, with the sampling bias correction proposed by Hardy & Vekemans (2002):

$$F_{ij} = \sum_{l=1}^{n_L} \left[\sum_{a=1}^{n_{a,l}} (p_{lai} - \bar{p}_{la})(p_{laj} - \bar{p}_{la}) + (N_l - 1)^{-1} \sum_{a=1}^{n_{a,l}} \bar{p}_{la}(1 - \bar{p}_{la}) \right] \Big/ \sum_{l=1}^{n_L} \sum_{a=1}^{n_{a,l}} \bar{p}_{la}(1 - \bar{p}_{la}) \qquad (1)$$

where n_L is the number of loci, $n_{a,l}$ is the total number of alleles at locus l, p_{lai} and p_{laj} are the frequencies of allele a at locus l in the relevant (as indicated in the third column of Table 1) haplotypes or genotypes of the i-th and j-th offspring, respectively, and $\bar{p}_{la}$ is the average frequency of allele a at locus l over all N_l genes with nonmissing information at locus l in the reference sample (see reference sample definitions in the last column of Table 1). Based on estimated parental correlations, PSA also derives effective numbers of fathers, mothers and parents contributing to each offspring patch (Table 1). Standard errors for parental correlation estimates are obtained via bootstrap resampling over individuals within-patches.

It is well known that genetic kinship coefficients suffer biases when the target individual pair and/or individuals that are close relatives of any of the individuals in the target pair are included in the sample used to compute reference allelic frequencies (Queller & Goodnight 1989). This may result, for instance, in negative estimates between offspring pairs that are less related than the average, potentially translating here into negative estimates of parental correlations among distant patches. PSA allows two (nonexclusive) optional corrections for this bias. The goal of both corrections is to account for all (or most) individuals who might be relatives of the target pair when defining the reference sample. First, provided that spatial information is available and that there is an observable decay of among-patch parental correlations with interpatch distance, it is possible to use this spatial information to define sets of unrelated gene pools. To do so, the user must set three threshold distances, defined as, respectively, the distance values above which the observed among-patch paternal, maternal and cross-parental correlations stabilize, which will be used by the program to identify a set of unrelated genotypes in the sample used for kinship coefficient calibration (see Robledo-Arnuncio *et al.* 2006 for more details). We refer here to this first approach as the 'threshold-distance correction'. Second, pairwise kinship coefficients F_{ij} can also be estimated with the 'leave-out' bias correction proposed by Queller & Goodnight (1989), whereby the $\bar{p}_{la}$ reference frequencies used in equation 1 are obtained from the data set excluding the patch(es) to which the target offspring pair (i, j) belongs, assuming that these patches will contain most of the individuals related to the target pair of individuals. The statistical properties of each of these two corrections are investigated below. We also initially tested an alternative leave-out correction involving simply the exclusion of the target offspring pair (not of the entire patch(es) to which they belong) in the computation of $\bar{p}_{la}$, but it yielded consistently inferior results that are not presented here.

Simulation analysis of method performance

We used Monte Carlo simulations to evaluate the expected relative bias (RBias) and relative accuracy (relative root mean square error, RRMSE) of parental correlation estimates obtained with PSA. We considered parental correlations as the analytical target here, noting that obtaining accurate estimates of the derivative effective parental numbers (which provide an intuitive reference of diversity) is a separate and nontrivial statistical exercise (Nielsen *et al.* 2003; Smouse & Robledo-Arnuncio 2005). We simulated n_P offspring patches sampled within a rectangular 40 × 40 central study area in a large (100 × 100) randomly distributed population of N cosexual parental plants. The pericarp and leaf genotypes of each of n_O simulated offspring per patch were available, using n_L loci with n_A equifrequent alleles each. Individual relative pollen ($\lambda_{p,i}$) and seed ($\lambda_{m,i}$) fecundity values for parental plants were assumed to be independently Poisson distributed, with mean λ. Pollen and seed dispersal from individual parents followed bivariate negative exponential kernels f_p and f_m, respectively, with means d_p and d_m, respectively. Within each simulated patch i, the expected proportion of offspring dispersed from maternal plant j was

$$\pi_{ij} = \frac{\lambda_{m,j} f_{m,ij}}{\sum_{q=1}^{N} \lambda_{m,q} f_{m,iq}}, \qquad (2)$$

where $f_{m,ij}$ is the probability given by the seed dispersal kernel of a seed travelling the distance that separates patch i from maternal plant j. In turn, the expected proportion of simulated offspring from maternal plant j sired by pollen donor k was

$$\tau_{jk} = \frac{\lambda_{p,k} f_{p,jk}}{\sum_{q=1}^{N} \lambda_{p,k} f_{p,jq}}, \qquad (3)$$

where $f_{p,jk}$ is the probability given by the pollen dispersal kernel of a pollen grain travelling the distance that separates maternal plant j from pollen donor k. Each independent simulation replicate comprised the following steps:

1 Distribute N parental plants randomly within the population and, for each of them, draw $\lambda_{p,i}$ and $\lambda_{m,i}$ independently from a Poisson distribution with mean λ.

2 Distribute n_P patches randomly within the central study area.

3 Generate n_O offspring for each patch i, with the number of offspring dispersed from each maternal plant drawn from a multinomial distribution with N classes with probabilities $\{\pi_{i1}, \pi_{i2}, \ldots, \pi_{iN}\}$ and n_O trials. Next, given the resulting number n_{ij} of simulated offspring from patch i dispersed from mother j, the number of them sired by each pollen donor was drawn from a multinomial distribution with N classes with probabilities $\{\tau_{j1}, \tau_{j2}, \ldots, \tau_{jN}\}$ and n_{ij} trials.

4 Generate two random alleles, from n_A possible classes, at each locus l for every parental individual in the population.

5 For every simulated offspring born to mother j and father k, generate a maternally inherited multilocus pericarp genotype identical to the multilocus genotype of mother j, and a biparentally inherited multilocus leaf genotype obtained by Mendelian segregation of alleles from mother j and father k at each locus l.

6 Estimate within- and among-patch parental correlations using PSA. Also compute the parametric (expected) parental correlations based on simulated parentage relationships.

The purpose of this simulation was not to recreate realistic seed and pollen dispersal processes, but rather to evaluate the performance of the estimators under different parental structure and sampling scenarios. A set of default parameters was chosen as reference, and the effect on the estimates of varying only one of them at a time was examined, considering the range of values described in the results. Default parameter values were as follows: $N = 2500$ potential parents, $n_P = 10$ patches sampled, $n_O = 20$ sampled offspring per patch, $n_L = 10$ loci, $n_A = 10$ alleles/locus, mean pollen dispersal distance $d_p = 5$, mean seed dispersal distance $d_m = 1$, and mean relative fecundity $\lambda = 10$. For each combination of parameters, the six simulation steps were repeated to generate $n_R = 1000$ independent realizations of the process, along with their associated parental correlation estimates, used to calculate expected estimation errors by comparing against their realized values in the simulated populations. In particular, we computed the expected RBias and the expected relative RMSE of the estimators of each kind of within-patch parental correlations (paternal, maternal or cross-parental, denoted generically as $\hat{Q}_w$) as follows:

$$\mathrm{RBias}(\hat{Q}_w) = \frac{1}{n_R n_P} \sum_{r=1}^{n_R} \sum_{i=1}^{n_p} \frac{\hat{Q}_{w,ri} - Q_w^*}{Q_w^*} \qquad (4)$$

$$\mathrm{RRMSE}(\hat{Q}_w) = \frac{1}{n_R n_P} \sum_{r=1}^{n_R} \sum_{i=1}^{n_P} \left(\frac{\hat{Q}_{w,ri} - Q_w^*}{Q_w^*} \right)^2 \qquad (5)$$

where $\hat{Q}_{w,ri}$ is the estimated parental correlation within the ith patch for the rth replicate data set, and Q_w^* is the mean of the actual within-patch parental correlation values realized in the simulations across all replicates. In the case of among-patch parental correlations, it would not be very informative to compute the average bias and average RMSE across all interpatch distance classes because, having variable levels of correlation, different distance classes may have absolute errors differing in orders of magnitude, while using relative errors would be problematic for long-distance classes (exhibiting small or null parental correlations), since they may take very large or even infinite values. Therefore, we chose to examine the distribution of among-patch parental correlation estimates against their expected values for different interpatch distance classes in a subset of the simulations.

Simulation results

Bias corrections

The 'threshold-distance' correction improved the accuracy (reduced the RRMSE) of all parental correlation estimates within and among patches, largely through bias reduction (Table 2; Fig. 1a,b). The improvement in within-patch correlation estimates was especially pronounced for paternal correlations and was observed under all three levels of parental structure considered, corresponding to maternal ($Q_{\mathrm{w}}^{\mathrm{m}}$), paternal ($Q_{\mathrm{w}}^{\mathrm{p}}$) and cross-parental ($Q_{\mathrm{w}}^{\mathrm{mp}}$) within-patch correlations ranging from 0.01 to 0.38 (Table 2). The precise choice of threshold-distance value did not have strong influence on the estimates, though slightly better accuracy was achieved for values corresponding to 2–3 times the average dispersal distance (Table 2). Relative estimation errors for $Q_{\mathrm{w}}^{\mathrm{m}}$ were smaller than those for $Q_{\mathrm{w}}^{\mathrm{mp}}$ and $Q_{\mathrm{w}}^{\mathrm{p}}$, as expected from the fact that $Q_{\mathrm{w}}^{\mathrm{m}}$ is directly computed from diploid pericarp genotypic pairs, while the calculation of $Q_{\mathrm{w}}^{\mathrm{mp}}$ and $Q_{\mathrm{w}}^{\mathrm{p}}$ is based on haplotypic pairs carrying half the amount of genetic information than diploid pericarps and requiring gametic phase inference. On the other hand, relative errors were higher for lower parental structure levels (Table 2).

At the interpatch level, the 'threshold-distance' approach corrected the systematic negative biases in the estimation of (near-zero or zero) parental correlation rates among distant patches, although residual negative biases remained in the case of (larger nonnull) parental correlations among closer patches (Fig. 1a,b). A combined application of the 'threshold-distance' and the 'leave-out' corrections together removed the residual biases at short interpatch distances, but at the cost of increased variance (Fig. 1c). Moreover, simultaneously applying the two corrections resulted in larger RBias and RRMSE of within-patch correlation estimates (compare Table S1, Supporting information versus Table 2). Finally, the 'leave-out' correction alone did not improve the accuracy of any of the estimates, neither within nor among patches (results not shown), so in the next two sections, we present only results obtained with either the 'threshold-distance' correction alone or simultaneously combined with the 'leave-out' approach.

Marker polymorphism

Increasing both the number of loci (n_{L}) and the number of alleles per locus (n_{A}) resulted in reduced RBias and RRMSE for all within-patch parental correlation estimates, especially for $Q_{\mathrm{w}}^{\mathrm{p}}$ and to a lesser degree for $Q_{\mathrm{w}}^{\mathrm{mp}}$ and $Q_{\mathrm{w}}^{\mathrm{m}}$ (Table 3). At constant total number of alleles across loci ($n_{\mathrm{A}} \cdot n_{\mathrm{L}}$), greater increase in $Q_{\mathrm{w}}^{\mathrm{p}}$ estimation

Table 2 Effect of the 'threshold-distance' correction and the level of parental structure on within-patch parental correlation estimates

			$\hat{Q}_w^m$			$\hat{Q}_w^p$			$\hat{Q}_w^{mp}$		
d_{m}	d_{p}	δ	Q^*	RBias	RRMSE	Q^*	RBias	RRMSE	Q^*	RBias	RRMSE
1	1	n_{c}	0.378	−0.085	0.179	0.314	−0.071	0.189	0.338	−0.117	0.198
		2		−0.004	0.169		0.008	0.189		−0.023	0.173
		3		−0.003	0.169		0.010	0.189		−0.022	0.173
		6		−0.002	0.169		0.011	0.188		−0.021	0.173
1	5	n_{c}	0.378	−0.082	0.185	0.028	−0.176	0.320	0.066	−0.247	0.340
		2		−0.001	0.177		−0.074	0.308		−0.022	0.257
		3		0.000	0.177		−0.066	0.309		−0.019	0.257
		6		0.001	0.177		−0.085	0.351		−0.027	0.280
5	5	n_{c}	0.027	−0.166	0.270	0.012	−0.211	0.477	0.015	−0.293	0.415
		2		−0.029	0.240		−0.121	0.495		−0.067	0.330
		3		−0.020	0.241		−0.098	0.499		−0.048	0.332
		6		−0.027	0.250		−0.094	0.528		−0.053	0.350

$\hat{Q}_w^m$, $\hat{Q}_w^p$ and $\hat{Q}_w^{mp}$, maternal, paternal and cross-parental within-patch correlation estimates, respectively; Q^*, actual mean parental correlation value realized in the simulations; RBias, relative bias; RRMSE, relative root mean square error; d_{m} and d_{p}, mean seed and pollen dispersal distance, respectively; δ, ratio between applied threshold distances and mean dispersal distances; n_{c}, no threshold-distance correction applied. Based on 1000 Monte Carlo replicates per row, assuming: $n_{\mathrm{P}} = 10$ patches sampled, $n_{\mathrm{O}} = 20$ offspring per patch sampled, $n_{\mathrm{L}} = 10$ loci, $n_{\mathrm{A}} = 10$ alleles/locus, population density = 0.25.

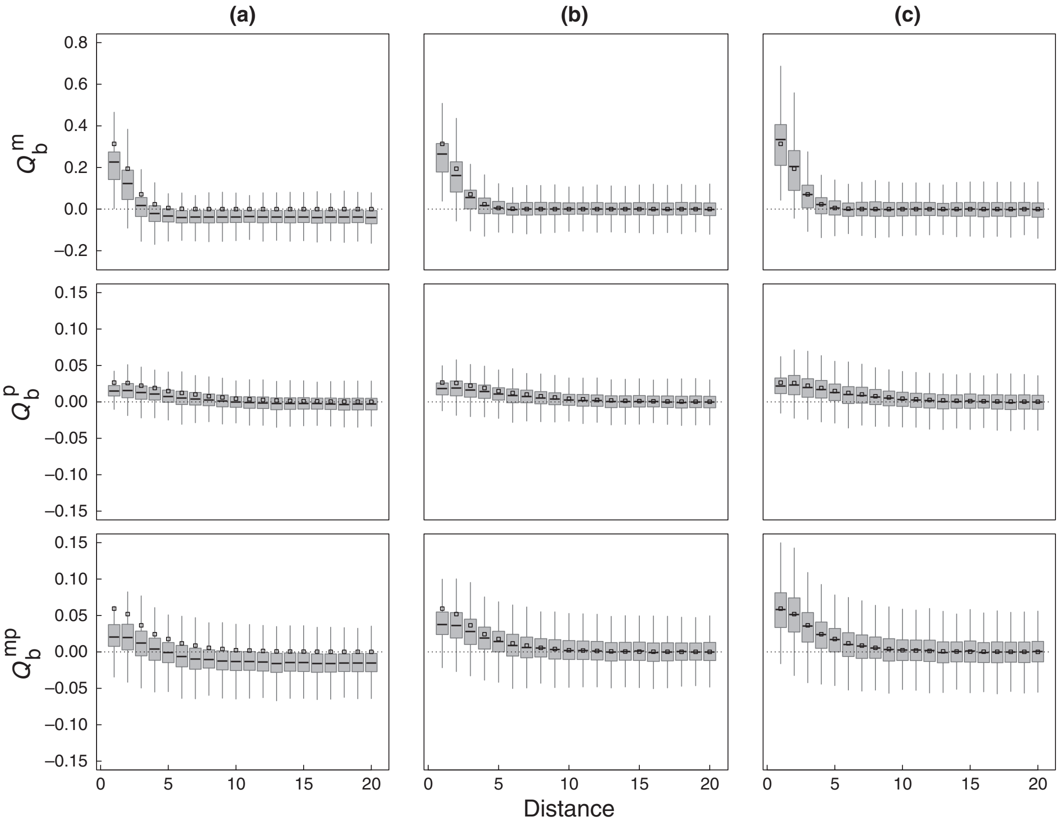

Fig. 1 Effect of the 'threshold-distance' and 'leave-out' corrections on among-patch parental correlation estimates. Boxplots show the distribution of maternal (Q_b^m), paternal (Q_b^p) and cross-parental (Q_b^{mp}) among-patch correlation estimates for different interpatch separation distance classes. Small squares indicate actual parental correlation values. Genetic kinship coefficients were as follows: (a) not corrected; (b) corrected using the 'threshold-distance' approach with threshold values set at twice the mean dispersal distance; or (c) corrected using both the 'threshold-distance' and the 'leave-out' approaches together, with threshold values set at twice the mean dispersal distance. Based on 1000 Monte Carlo replicates assuming: $n_P = 10$ patches sampled, $n_O = 20$ offspring per patch sampled, $n_L = 10$ loci, $n_A = 10$ alleles/locus, mean seed and pollen dispersal distances $d_m = 1$ and $d_m = 5$, respectively, and population density = 0.25.

Table 3 Effect of marker polymorphism on within-patch parental correlation estimates (actual correlation values in parentheses)

		$\hat{Q}_w^m$ (0.378)		$\hat{Q}_w^p$ (0.028)		$\hat{Q}_w^{mp}$ (0.066)	
n_L	n_A	RBias	RRMSE	RBias	RRMSE	RBias	RRMSE
5	5	0.004	0.202	−0.137	0.469	−0.022	0.329
5	10	0.003	0.179	−0.091	0.369	−0.018	0.285
5	20	0.002	0.171	−0.047	0.299	−0.010	0.252
10	5	0.001	0.183	−0.140	0.407	−0.024	0.289
20	5	0.000	0.168	−0.147	0.328	−0.023	0.251
10	10	−0.001	0.177	−0.074	0.308	−0.022	0.257
20	20	−0.0014	0.175	−0.040	0.239	−0.015	0.227
30	30	−0.001	0.162	−0.023	0.226	−0.014	0.215
50	2	−0.001	0.172	−0.362	0.460	−0.065	0.269

$\hat{Q}_w^m$, $\hat{Q}_w^p$ and $\hat{Q}_w^{mp}$, maternal, paternal and cross-parental within-patch correlation estimates, respectively; RBias, relative bias; RRMSE, relative root mean square error; n_L, number of loci; n_A, number of alleles per locus. Based on 1000 Monte Carlo replicates per row, assuming: $n_P = 10$ patches sampled, $n_O = 20$ offspring per patch sampled, mean seed and pollen dispersal distances $d_m = 1$ and $d_m = 5$, respectively, population density = 0.25 and threshold-distance correction applied with threshold values set at twice the mean dispersal distance.

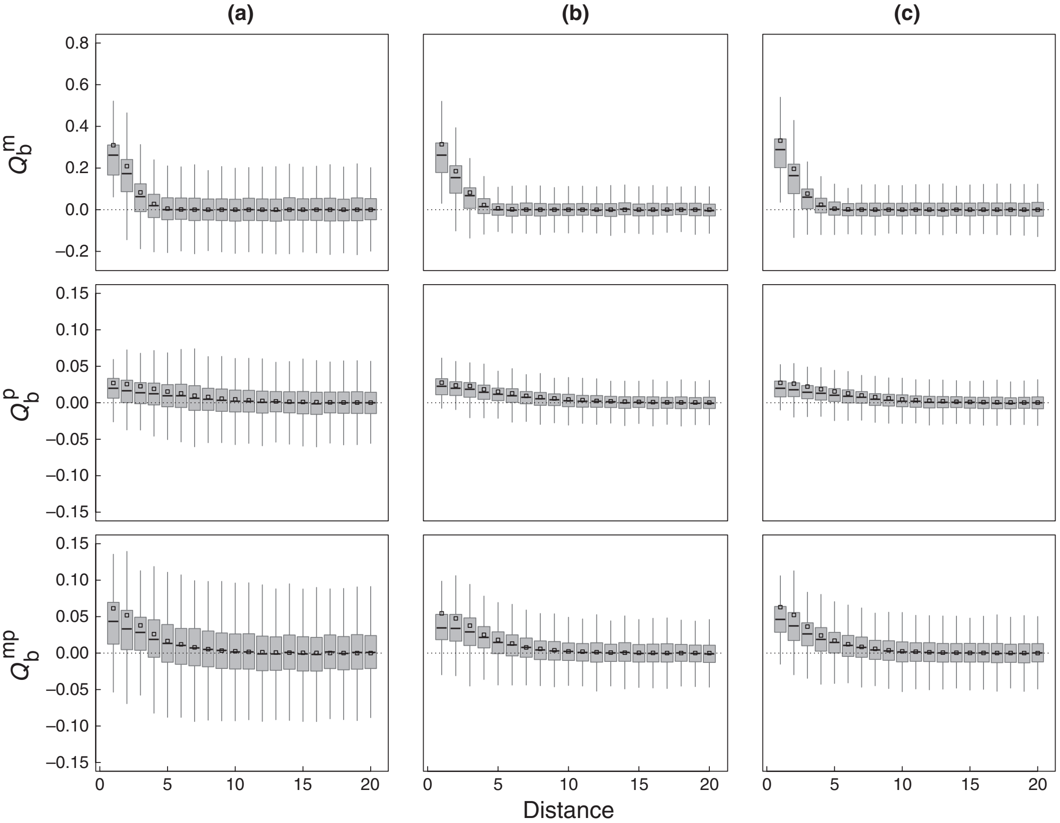

Fig. 2 Effect of marker polymorphism on among-patch parental correlation estimates. Boxplots as in Fig. 1. The assumed number of loci (n_L) and number of alleles per locus (n_A) were as follows: (a) $n_L = 5$ and $n_A = 5$; (b) $n_L = 5$ and $n_A = 20$; or (c) $n_L = 20$ and $n_A = 5$. Based on 1000 Monte Carlo replicates assuming: $n_P = 10$ patches sampled, $n_O = 20$ offspring per patch sampled, mean seed and pollen dispersal distances $d_m = 1$ and $d_m = 5$, respectively, population density = 0.25 and threshold-distance correction applied with threshold values set at twice the mean dispersal distance.

accuracy was achieved using a smaller number of more polymorphic loci than using a larger number of less polymorphic loci. The RBias of Q_w^p increased sharply when using biallelic loci (but not that of Q_w^m; Table 3), suggesting that the sensitivity of Q_w^p to decreasing numbers of alleles per locus derives from greater uncertainty in resolving paternal genotypes. Among-patch parental correlation estimates benefited from increased genetic resolution as well, exhibiting the same sensitivity to the number of alleles and loci as the within-patch correlation estimates (Fig. 2). Similar marker polymorphism effects were encountered when using the 'threshold-distance' and 'leave-out' corrections together (Table S2, Fig. S1, Supporting information). For all levels of marker polymorphism considered, the combined correction strategy resulted in consistently higher RBias and RRMSE for within-patch correlation estimates, and lower RBias but higher RRMSE for among-patch correlation estimates, than did the threshold-distance correction alone (Table 3 vs. Table S2, Supporting information and Fig. 2 vs. Fig. S1, Supporting information).

Sampling intensity

Increasing the total number of sampled offspring reduced both the RBias and RRMSE of all three within-patch parental correlation estimates, especially those of Q_w^p and Q_w^{mp} (Table 4). Assuming a fixed total sample size, sampling more offspring per patch was more beneficial than sampling more patches, in terms of reducing Q_w^p estimation errors, while the reverse was true for Q_w^m and Q_w^{mp} (Table 4). More total sampled offspring also decreased the bias and specially the variance of among-patch parental correlation estimates (Fig. 1a vs. Fig. 1b,c). While increasing the number of offspring sampled per patch was more efficient at reducing the variance of Q_b^p, Q_b^m and Q_b^{mp}, sampling more patches was more effective at reducing their RBias (Fig. 3b,c). Finally, under all sampling intensity scenarios considered, adding the 'leave-out' correction increased the bias and variance of within-patch parental correlation estimates (Table S3, Supporting information vs. Table 4), but it reduced the bias of among-patch estimates at the cost of increased variance (Fig. S2, Supporting information vs. Fig. 3).

Table 4 Effect of sampling intensity on within-patch parental correlation estimates (actual correlation values in parentheses)

		$\hat{Q}_w^m$ (0.378)		$\hat{Q}_w^p$ (0.028)		$\hat{Q}_w^{mp}$ (0.066)	
n_P	n_O	RBias	RRMSE	RBias	RRMSE	RBias	RRMSE
10	5	0.008	0.230	−0.048	0.953	−0.003	0.505
10	10	0.004	0.195	−0.077	0.525	−0.015	0.345
10	20	−0.001	0.177	−0.074	0.308	−0.022	0.257
10	40	−0.004	0.154	−0.089	0.207	−0.018	0.194
10	80	−0.001	0.161	−0.091	0.166	−0.019	0.172
2	20	0.011	0.395	−0.011	0.817	0.004	0.600
5	20	0.002	0.247	−0.070	0.445	−0.016	0.357
20	20	−0.001	0.119	−0.072	0.220	−0.015	0.172
40	20	−0.001	0.089	−0.086	0.170	−0.012	0.131

$\hat{Q}_w^m$, $\hat{Q}_w^p$ and $\hat{Q}_w^{mp}$, maternal, paternal and cross-parental within-patch correlation estimates, respectively; RBias, relative bias; RRMSE, relative root mean square error; n_P, number of sampled patches; n_O, number of sampled offspring pet patch. Based on 1000 Monte Carlo replicates per row, assuming: n_L = 10 loci, n_A = 10 alleles/locus, mean seed and pollen dispersal distances d_m = 1 and d_m = 5, respectively, population density = 0.25 and threshold-distance correction applied with threshold values set at twice the mean dispersal distance.

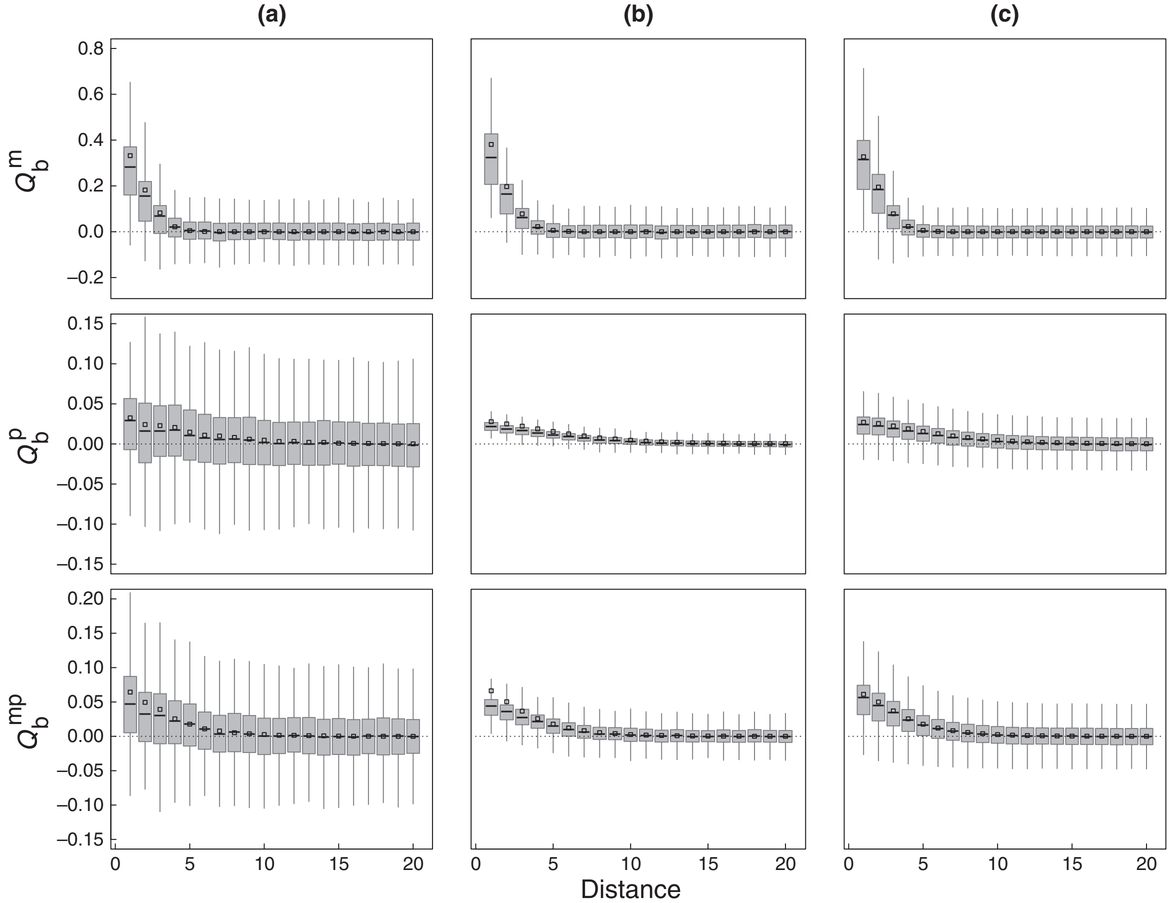

Fig. 3 Effect of sampling intensity on among-patch parental correlation estimates. Boxplots as in Fig. 1. The assumed number of sampled patches (n_P) and number of sampled offspring per patch (n_O) were as follows: (a) n_P = 10 and n_O = 5; (b) n_P = 10 and n_O = 80; or (c) n_P = 40 and n_O = 20. Based on 1000 Monte Carlo replicates assuming: n_L = 10 loci, n_A = 10 alleles per locus, mean seed and pollen dispersal distances d_m = 1 and d_m = 5, respectively, population density = 0.25, and threshold-distance correction applied with threshold values set at twice the mean dispersal distance.

Sampling recommendations

Results of the simulation study suggest that PSA can provide fairly accurate parental correlation estimates from affordable sampling effort. Larger sample sizes will be necessary to achieve a given level of *relative* accuracy for weaker parental structures, which should be expected for higher parental densities and longer-range pollen and seed dispersal. As a rule of thumb, it is desirable to sample a total of at least 200 offspring, with no <10–20 per patch. If within-patch parental correlations are the main interest, sampling more offspring per patch will be most efficient in reducing estimation errors, while larger numbers of patches should be favoured when the distribution of pairwise among-patch correlations is a central goal. In any case, it is generally advisable to use spatial information to calibrate kinship coefficients (apply the 'threshold-distance' correction option), thereby minimizing the RRMSE of both within- and among-patch correlation estimates. This correction requires sampling some patch pairs far enough apart to ensure null or near-null pairwise parental correlations among them, permitting thus designation of meaningful threshold distances, beyond which gene pools are unrelated. The 'leave-out' correction should only be used in conjunction with the 'threshold-distance' correction, and only if the accuracy of among-patch parental correlation estimates is critical and sample sizes are rather large. Finally, 5–10 polymorphic loci should provide reasonably accurate parental correlation estimates, but larger numbers of highly polymorphic loci will help minimizing estimation errors.

Applications

Ecologists, conservation biologists and ecosystem managers are concerned with contemporary gene flow as an important source of genetic variability and outcrossing, especially for fragmented and low density populations (Ledig 1992; Ellstrand & Elam 1993; Sork *et al.* 1999). By providing estimates of parental correlations and effective numbers of parents contributing to plant offspring groups, the PSA software is useful to quantify the impact of different demographic scenarios on contemporary reproductive and dispersal processes mediating gene flow. Effective parental numbers provide in particular a measure of the potential for genetic drift, inbreeding and kin competition on a local scale. In addition, the decay in among-patch parental correlations with distance reflects the degree of isolation among patches as determined by the spatial range of pollen and seed dispersal. Decoupling pollen and seed dispersal is interesting because it allows the inference of the relative contributions of different dispersal vectors to the genetic structure of natural regeneration. For instance, biologists may want to assess whether the loss of a pollinator might be mitigated by the seed dispersal vector, in terms of the total effective number of parents contributing to offspring groups (e.g. Wilcock & Neiland 2002; Lowe *et al.* 2005). Conversely, many habitats are losing their seed dispersal agents (e.g. Howe & Miriti 2004; Wang *et al.* 2007), and it would be useful to know the extent to which effective pollen dispersal might mitigate the impacts of reduced seed dispersal on effective parental numbers. Being able to assess the differential contribution of contemporary pollen and seed flow allows understanding their respective role in shaping genetic and demographic patterns across the landscape and across time, especially under current environmental perturbations.

Program availability

A program written in C++ language that implements the estimation method described in this paper, with user manual and test data files, is freely available at https://sites.google.com/site/jjrobledo2/software.

Acknowledgements

This work was supported by Ramón y Cajal research fellowships to JJRA and DG and by CGL2009-09428 project from the Spanish Ministry of Economy and Competitiveness. PES was supported by the New Jersey Agricultural Experiment Station and the US Department of Agriculture (NJAES/USDA-17111) and the National Science Foundation (NSF-DEB-0514956). VLS and DG were supported by UCLA and the National Science Foundation (NSF-DEB-0516529).

References

Baskin CC, Baskin JM (2000) *Seeds: Ecology, Biogeography and Evolution of Dormancy and Germination*. Academic Press Inc., San Diego.

Blouin MS (2003) DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends in Ecology and Evolution*, **18**, 503–511.

Burczyk J, Adams WT, Birkes DS, Chybicki IJ (2006) Using genetic markers to directly estimate gene flow and reproductive success parameters in plants on the basis of naturally regenerated seedlings. *Genetics*, **173**, 363–372.

Chybicki I, Burczyk J (2010) NM+: software implementing parentage-based models for estimating gene dispersal and mating patterns in plants. *Molecular Ecology Resources*, **10**, 1071–1075.

Ellstrand NC, Elam DR (1993) Population genetics of small population size: implications for plant conservation. *Annual Review of Ecology and Systematics*, **23**, 217–242.

Fenner M (1992) *Seeds: The Ecology of Regeneration in Plant Communities*. CAB International, Wallingford.

Gerber S, Chabrier P, Kremer A (2003) FaMoz: a software for parentage analysis using dominant, codominant and uniparentally inherited markers. *Molecular Ecology Notes*, **3**, 479–481.

Godoy JA, Jordano P (2001) Seed dispersal by animals: exact identification of source trees with endocarp DNA microsatellites. *Molecular Ecology*, **10**, 2275–2283.

Goto S, Shimatani K, Yoshimaru H, Takahashi Y (2006) Fat-tailed gene flow in the dioecious canopy tree species *Fraxinus mandshurica* var. *japonica* revealed by microsatellites. *Molecular Ecology*, **15**, 2985–2996.

Grivet D, Smouse PE, Sork VL (2005) A novel approach to an old problem: tracking dispersed seeds. *Molecular Ecology*, **14**, 3585–3595.

Grivet D, Robledo-Arnuncio JJ, Smouse PE, Sork VL (2009) Relative contribution of contemporary pollen and seed dispersal to the effective parental size of seedling populations of Californian valley oak (*Quercus lobata*, Née). *Molecular Ecology*, **18**, 3967–3979.

Hardy OJ, Vekemans X (2002) SPAGEDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes*, **2**, 618–620.

Howe GT, Miriti MN (2004) When seed dispersal matters. *BioScience*, **54**, 651–660.

Ibarra-Manríquez G, Martínez Ramos M, Oyama K (2001) Seedling functional types in a lowland rain forest in Mexico. *American Journal of Botany*, **88**, 1801–1812.

Irwin AJ, Hamrick JL, Godt MJW, Smouse PE (2003) A multiyear estimate of the effective pollen donor pool for *Albizia julibrissin*. *Heredity*, **90**, 187–194.

Iwaizumi MG, Watanabe A, Ubukata M (2007) Use of different seed tissues for separate biparentage identification of dispersed seeds in conifers: confirmations and practices for gene flow in *Pinus densiflora*. *Canadian Journal of Forest Research*, **37**, 2022–2030.

Jones OR, Wang J (2010) COLONY: aprogram for parentage and sibship inference from multilocus genotype data. *Molecular Ecology Resources*, **10**, 551–555.

Kalinowski ST, Wagner AP, Taper ML (2006) ML-RELATE: a computer program for maximum likelihood estimation of relatedness and relationship. *Molecular Ecology Resources*, **6**, 575–579.

Karubian J, Sork VL, Roorda T, Duraes R, Smith TB (2010) Destination-based seed dispersal homogenizes genetic structure of a tropical palm. *Molecular Ecology*, **19**, 1745–1753.

Ledig FT (1992) Human impacts on genetic diversity in forest ecosystems. *Oikos*, **63**, 87–108.

Loiselle BA, Sork VL, Nason J, Graham C (1995) Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *American Journal of Botany*, **82**, 1420–1425.

Lowe AJ, Boshier D, Ward M, Bacles CFE, Navarro C (2005) Genetic resource impacts of habitat loss and degradation; reconciling empirical evidence and predicted theory for neotropical trees. *Heredity*, **95**, 255–273.

Marshall TC, Slate J, Kruuk LEB, Pemberton JM (1998) Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular Ecology*, **7**, 639–655.

Nielsen R, Tarpy DR, Reeve K (2003) Estimating effective paternity number in social insects and the effective number of alleles in a population. *Molecular Ecology*, **12**, 3157–3164.

Queller DC, Goodnight KF (1989) Estimating relatedness using genetic markers. *Evolution*, **43**, 258–275.

Robledo-Arnuncio JJ, Austerlitz F, Smouse PE (2006) A new method of estimating the pollen dispersal curve independently of effective density. *Genetics*, **173**, 1033–1045.

Rousset F (2008) GENEPOP'007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources*, **8**, 103–106.

Rousset F, Leblois R (2012) Likelihood-based inferences under isolation by distance: two-dimensional habitats and confidence intervals. *Molecular Biology and Evolution*, **29**, 957–973.

Smouse PE, Robledo-Arnuncio JJ (2005) Measuring the genetic structure of the pollen pool as the probability of paternal identity. *Heredity*, **94**, 640–649.

Smouse PE, Sork VL, Scofield DG, Grivet D (2012) Using seedling and pericarp tissues to determine maternal parentage of dispersed Valley oak recruits. *Journal of Heredity*, **103**, 250–259.

Sork VL, Nason J, Campbell DR, Fernandez JF (1999) Landscape approaches to historical and contemporary gene flow in plants. *Trends in Ecology & Evolution*, **14**, 219–224.

Thomas PA (2000) *Trees: Their Natural History*. Cambridge University Press, Cambridge.

Vander Wall SB (2001) The evolutionary ecology of nut dispersal. *The Botanical Review*, **67**, 74–117.

Wang J (2011) COANCESTRY: a program for simulating, estimating and analysing relatedness and inbreeding coefficients. *Molecular Ecology Resources*, **11**, 141–145.

Wang BC, Sork VL, Leong MT, Smith TB (2007) Hunting of mammals reduces seed removal and dispersal of the afrotropical tree *Antrocaryon klaineanum* (Anacardiaceae). *Biotropica*, **39**, 340–347.

Wilcock C, Neiland R (2002) Pollination failure in plants: why it happens and when it matters. *Trends in Plant Science*, **7**, 270–277.

Ziegenhagen B, Liepelt S, Kuhlenkamp V, Fladung M (2003) Molecular identification of individual oak and fir trees from maternal tissues of their fruits or seeds. *Trees Structure and Function*, **17**, 345–350.

The presented software implements analytical methods conceived and developed by J.J.R.A. and P.E.S. J.J.R.A. coded the program, conducted the numerical study and wrote the first draft of the paper. D.G., V.L.S. and P.E.S. contributed to the development of the final version of the paper.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1 Effect of the combined 'threshold-distance' and 'leave-out' corrections and the level of parental structure on within-patch parental correlation estimates.

Table S2 Effect of marker polymorphism on within-patch parental correlation estimates (actual correlation values in parentheses) when applying the 'threshold-distance' and the 'leave-out' corrections simultaneously.

Table S3 Effect of sampling intensity on within-patch parental correlation estimates (actual correlation values in parentheses) when applying the 'threshold-distance' and the 'leave-out' corrections simultaneously.

Fig. S1 Effect of marker polymorphism on among-patch parental correlation estimates when applying the 'threshold-distance' and 'leave-out' corrections together.

Fig. S2 Effect of sampling intensity on among-patch parental correlation estimates when applying the 'threshold-distance' and 'leave-out' corrections together.

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