Heredity (2003) 90, 187–194
© 2003 Nature Publishing Group All rights reserved 0018-067X/03 $25.00
www.nature.com/hdy

A multiyear estimate of the effective pollen donor pool for *Albizia julibrissin*

AJ Irwin[1,2], JL Hamrick[3,4], MJW Godt[3] and PE Smouse[1]

[1]*Department of Ecology, Evolution & Natural Resources, Rutgers University, New Brunswick, NJ 08901, USA;* [2]*Institute of Marine and Coastal Sciences, Rutgers University, New Brunswick, NJ 08901, USA;* [3]*Department of Plant Biology, University of Georgia, Athens, GA 30602, USA;* [4]*Department of Genetics, University of Georgia, Athens, GA 30602, USA*

Studies of pollen movement in plant populations are often limited to a single reproductive event, despite concerns about the adequacy of single-year measures for perennial organisms. In this study, we estimate the effective number of pollen donors per tree from a multiyear study of *Albizia julibrissin* Durazz (mimosa, Fabaceae), an outcrossing, insect-pollinated tree. We determined 40 seedling genotypes for each of 15 seed trees during 4 successive years. A molecular analysis of variance of the pollen gametes fertilizing the sampled seeds was used to partition variation in pollen pools among seed trees, among years, and within single tree-year collections. Using these variance components, we demonstrate significant male gametic variability among years for individual trees. However, results indicate that yearly variation in the 'global pollen pool', averaged over all 15 seed trees for these 4 years, is effectively zero. We estimate the effective number of pollen donors for a single mimosa tree (N_{ep}) to be 2.87. Single season analyses yield $N_{ep} \sim 2.05$, which is 40% less than the value of N_{ep} estimated from 4 years of data. We discuss optimal sampling for future studies designed to estimate N_{ep}. Studies should include more trees, each sampled over at least a few years, with fewer seeds per tree per year than are needed for a traditional parentage study.

Heredity (2003) **90,** 187–194. doi:10.1038/sj.hdy.6800215

Keywords: effective pollen pool size; *Albizia julibrissin*; temporal variation; sampling strategy; TwoGener

Introduction

Currently, there is considerable interest in describing mating patterns for plant species. Much of this interest stems from the potential impact of anthropogenic fragmentation on the breeding structure of natural plant populations (eg, see Sork *et al*, 1998). Most studies of plant breeding patterns are single-season studies (but see Schoen and Stewart, 1986; Hamrick and Murawski, 1990; Godt and Hamrick, 1993; Schnabel and Hamrick, 1995). For perennial organisms, however, single-season studies may provide a limited characterization of generational patterns of mating. The effective pollen donor pool size per generation is of interest, since annual variation in pollen composition received by an individual is likely to increase genetic variation among progeny. This may have long-term effects on the ability of populations to maintain effective population sizes that can withstand demographic and genetic stochasticity (Schemske *et al*, 1994).

Individual perennial plants often do not flower every year and environmental effects (eg variation in pollinators or their abundance or climatic variation) may generate substantial interannual variation in the quantity and genetic quality of pollen received by a particular plant. Temporal variability in the pollen donors of particular plants or among plants within populations is poorly understood, but it is likely that there is year-to-year variation that is not captured by single-season studies. The importance of such variation to the overall genetic composition of the population may be slight if progeny recruitment is characterized by the establishment of a few progeny from each reproductive event. In that case, temporal heterogeneity in the genetic composition of pollen and progeny may be smoothed over an individual's lifetime reproduction. In contrast, if successful recruitment is episodic, with only a very few successful recruitment events during the reproductive lifetime of an individual, temporal variation in breeding patterns and pollen composition could have a marked effect on the genetic composition of the resulting progeny.

Paternity procedures are often used to identify individuals that have contributed a particular male gamete to a seed (eg Ellstrand and Marshall, 1985; Meagher, 1986; Schoen and Stewart, 1986; Broyles and Wyatt, 1990; Devlin and Ellstrand, 1990). However, genetically characterizing and identifying all potential fathers is a difficult task in large or continuous populations, since some pollen may originate from nonidentifiable (presumably distant and unsampled) pollen donors. Fortunately, estimates of the effective number of pollen donors (N_{ep}) per plant can be obtained without identifying the precise father of each seedling. We have previously developed methods to estimate the effective number of pollen donors per plant for a single (yearly) fertilization event (Sork *et al*, 1998; Austerlitz and Smouse, 2001; Smouse *et al*, 2001). Genotypes of seed plants and their progeny provide data for this analysis.

Correspondence: Andrew J Irwin, Institute of Marine and Coastal Sciences, Rutgers University, 71 Dudley Road, New Brunswick, NJ 08901-8521, USA. E-mail: ajirwin@rci.Rutgers.Edu

Received 8 October 2001; accepted 7 October 2002

We introduced a new statistic (Φ_{ft}), extractable from a molecular analysis of variance (AMOVA) of the male gametes that fertilize several individuals. The Φ_{ft} statistic is the intraclass correlation coefficient, describing the proportion of the pollen pool (male gametic) variance that is attributable to differentiation among pollen pools sampled by designated maternal plants. This statistic is analogous to Wright's (1951) F_{st} coefficient of differentiation among populations. Here, the strata are individual plants and the within-stratum replication is provided by multiple progeny genotypes. The effective number of pollen donors, N_{ep}, for individual plants can be deduced from an estimate of Φ_{ft}, given simple pollen-dispersal models (Austerlitz and Smouse, 2001; Smouse *et al*, 2001). The question of how well a single-season study characterizes the pollen-donor pool over an entire generation motivates this study.

Objectives

Year-to-year variation in pollen production and pollinator behaviors suggests that an individual may sample pollen from a changing array of pollen donors over time. Regardless of whether year-to-year changes in the effective male gametic pool are large or small, year-to-year variation in pollen donors can only expand the total number (and probably the spatial extent) of pollen donors for any long-lived perennial, relative to the number (and spacing) of pollen donors sampled during a single reproductive event. Studies are needed to assess how well (indeed, whether) a single season's study of pollen movement can be extrapolated to a per-generation inference. In this study, we address the following questions:

- Are yearly pollen donors for a given plant simply random samples from a temporally stable pollen donor pool for that plant, or is there significant year-to-year variation in pollen donor pools?
- Is year-to-year variation in pollen donors best viewed as replication error within a plant, or is there meaningful year-to-year variation in the global pollen pool, averaged over plants within a population?
- What is the effective overlap of pollen profiles for single plants across time, and how does that affect the effective number of pollen donors, N_{ep}, per plant over several reproductive events?
- How much larger is a multiyear estimate of the effective pollen donor pool compared to the single-year estimate per plant?

Mimosa study

This is a multiyear study of successful fertilizations in naturalized populations of mimosa (*Albizia julibrissin*). The study was designed to gauge the influence of year-to-year variation in the array of pollen donors on the effective size of the paternal donor pool per seed tree. Mimosa is a small (<15 m) mimosoid legume whose native range extends from Iran to China and Japan. An attractive and profusely flowering tree, mimosa was introduced into the southern US as an ornamental at least 150 years ago (Elias, 1987). With a range extending from Maryland south to Florida, and west to Texas and California, mimosa is now naturalized throughout the southern US (Elias, 1987). Mimosa prospers in full sun, and naturalized populations are often found along roadsides, around abandoned home sites and along forest edges. Mimosa is typically found in small populations of three to 25 flowering trees, but larger populations (100+ flowering individuals) and isolated trees are not uncommon. In northeastern Georgia, mimosa flowers from early May to August. Flowers occur in white, pink or reddish inflorescences clustered at the ends of terminal and lateral branches, and they are visited by hummingbirds and a wide array of insects. Each inflorescence produces one to five fruits (pods). Since pollen is dispersed as a cluster (a polyad), seeds within the majority of pods are full-sibs (ie offspring of a single pollen donor; Hamrick unpublished data). Large, sunlit trees produce copious numbers of flowers and fruits annually. In open habitats, trees initiate flowering at small sizes (<2.5 m). Seeds are produced in thin, flat pods that are readily dispersed by wind, (Hamrick and Godt, personal observation). Since mimosa populations are colonized, initiate flowering, become mature and senesce within 10–20 years, annual changes in the breeding structure of their populations can be tracked over a substantial proportion of a population's lifetime, an unusual feature for a tree species.

Field sampling

The data we present are a subset from a long-term study of mimosa undertaken in Athens, GA. To ensure a completely balanced design, we selected 15 trees from the data set, each of which provided large numbers of genetically assayed progeny in 1989, 1990, 1991 and 1992. 'Maternal' trees were selected to cover a range of intertree distances, from a few meters to 7 km. Approximately 60 fruits (with 6–12 seeds per pod) were collected from each tree during 4 successive years. Seeds from each tree were bulked for each year and 100 seeds per tree per year were planted. A random sample of 60 seedlings per tree was assayed for eight polymorphic allozyme loci. Seedlings with incomplete genotypic typings were removed from the data set. The 15 sibships were subsampled to obtain 40 random progeny per year, per tree (40 seedlings × 15 trees × 4 years = 2400 seedlings). More seedlings could have been used, but this would have resulted in uneven sampling of mothers and years, and would have made variance component estimation more complicated. With 2400 seedlings, we have ample replication for the study; additional seedlings would have principally improved estimates of the residual variation, rather than estimates of 'among-trees' and 'among-years' components.

The existence of full-sib progeny within each pod raises the possibility that our sampling scheme could include progeny originating from the same fruit. However, the selection of 40 progeny from a bulked sample of seeds extracted from 60 pods (≈600 seeds) makes it unlikely that there are more than a few same-fruit progeny pairs. The small number of same-fruit progeny should have minimal impact on our estimates of the Φ-statistics.

Genetic characterization

Genotypes for eight polymorphic allozyme loci were obtained; each locus provided clear genetic resolution that was reliable from gel-to-gel and year-to-year. The loci analyzed were cathodal peroxidase (*cPer*, E.C. 1.11.1.7), fluorescent esterase (*Fe-2*, E.C. 3.1.1-), phosphoglucomutase (*Pgm-2*, E.C. 5.4.2.2), diaphorase (*Dia-1*, E.C. 1.6.99-), isocitrate dehydrogenase (*Idh-1*, E.C. 1.1.1.42), phosphoglucoisomerase (*Pgi-2*, E.C. 5.3.1.9), aspartate aminotransferase (*Aat-1*, E.C. 2.6.1.1) and 6-phosphogluconate dehydrogenase (*6Pgdh*, E.C. 1.1.1.44). Greater Athens' global frequencies for each locus are shown in Table 1. Seed trees were not genotyped, but with 160 seedlings per tree, their genotypes were obvious and were consistent across years. The multilocus exclusion probability based on these eight loci ($E_L \sim 0.76$; Chakravarti and Li, 1983; Jamieson, 1994) was not sufficient to provide strong inference on precise male parentage for individual seedlings, but we have shown elsewhere (Smouse *et al*, 2001) that it is more than sufficient for the TwoGener analyses described here.

Data analysis

Characterization of male gametes

Since seedlings are sampled from maternal trees whose genotypes have been inferred, pollen genotypes can be deduced by subtracting the maternal gametic combination. This is straightforward, except when the mother and seedling share the same heterozygous genotype (say G_kG_l). In that case, the pollen contribution to the seedling is ambiguous. At some of the loci in this study, as many as 25–50% of the pollen contributions are ambiguous (exact numbers of unambiguous alleles at each locus are given in Table 1a). We have shown how to deal with this situation (Smouse *et al*, 2001), using likelihood-based parentage analysis techniques, but to do that, we require estimates of allele frequencies for polymorphic loci in the pollen pools of individual trees, probably for each year of the study. The usual strategy is to use the unambiguous male gametes of a particular plant to establish relevant allele frequencies, while ignoring ambiguous gametes. The difficulty is that by ignoring ambiguous male gametes, we underestimate the allele frequencies in question and overestimate frequencies of unambiguous alleles at the locus. Since ambiguous heterozygotes most often involve the most common (polymorphic) alleles at a locus, this procedure systematically underestimates frequencies of common alleles and overestimates those of rare alleles. We can correct this bias by using allelic counts of the unambiguous gametes to 'fill the holes', basically adapting a missing value technique to provide gametic estimates for the ambiguous seedlings, and then computing adjusted allele frequencies from all seedlings (see Appendix A). With such adjustments, we achieve bias-free allele frequency estimates and ensure full sample sizes, $N_{ij} = 40$ seedlings, for each locus and each seed tree in each year. Entries in Table 1a are biased estimates from the unambiguous gametes; those in Table 1b are unbiased estimates from the correction procedure described above and elaborated in Appendix A. The differences are not large, and the exclusion probability is the same (within 0.2%) for both estimates.

Choice of pollen pools

There are several possible sources of unambiguous pollen frequencies that could be employed, but the best

Table 1 Global allele frequencies and sample sizes for the Athens pollen pool, extracted from unambiguous male gametes of *Albizia julibrissin*, for each of eight allozyme loci

Allele	cPer-1	Fe-2	Pgm-2	Dia-1	Idh-1	Pgi-2	Aat-1	6Pgdh
(a) *Estimates from unambiguous pollen counts*								
2	—	—	—	0.0005	—	0.247	—	—
3	—	0.823	0.003	0.065	—	0.261	0.017	0.003
4	0.666	0.177	0.787	0.932	0.842	0.325	0.982	0.942
5	0.334	—	0.207	0.003	—	0.001	0.0004	0.054
6	—	—	0.003	—	0.158	0.147	—	—
7	—	—	—	—	—	0.018	—	—
Sample size (*N*)	1170	1591	1749	2160	2053	1879	2308	2173
Excl. probability	0.1730	0.1246	0.1444	0.0605	0.1152	0.5014	0.0172	0.0520
$E_L = 1 - \prod_{l=1}^{L=8} (1 - E_l) = 0.7608 \approx 0.76$								
(b) *Estimates from total pollen counts, using tree- and year-specific averages to estimate ambiguous gametic vectors (N=2400)*								
2	—	—	—	0.0004	—	0.249	—	—
3	—	0.840	0.002	0.081	—	0.239	0.020	0.003
4	0.638	0.160	0.751	0.916	0.859	0.359	0.980	0.938
5	0.362	—	0.245	0.003	—	0.001	0.0004	0.059
6	—	—	0.002	—	0.141	0.138	—	—
7	—	—	—	—	—	0.015	—	—
Excl. probability	0.1776	0.1164	0.1562	0.0725	0.1063	0.4894	0.0194	0.0557
$E_L = 1 - \prod_{l=1}^{L=8} (1 - E_l) = 0.7597 \approx 0.76$								

Exclusion probabilities (E) are given for each locus and over all loci for (a) and (b).

choice depends on how pollen donor variation is distributed over space and time, something that is unknown, a priori. There are at least 4 choices: (a) estimates from the 40 offspring in each seed tree-year combination, (b) estimates from all 160 offspring of a given tree, effectively averaged over the 4 years of the study, (c) estimates from all 600 offspring from a particular year, effectively averaged over all 15 trees or (d) global estimates obtained from all 2400 seedlings, effectively averaged over all 15 trees and all 4 years of the study. Estimates based on pooled samples will inevitably reduce subsequent measures of variability among those pooled samples, which will tend to favor the null hypothesis of no stratification. In other words, the choice of strategy affects the answers obtained to some degree.

Analysis of Molecular Variance (AMOVA)

Excoffier *et al* (1992) describe the AMOVA as an analysis of variance based on a matrix of genetic distances between individuals. Smouse *et al* (2001) extended Excoffier's treatment to haploid pollen genotypes, dubbing the analytical procedure TwoGener. In this study, we analyze 2400 offspring. A crossclassified analysis, with trees and years as main effects, has 60 strata, with 40 replicate observations (male gametes) per stratum. A 2400 × 2400 distance matrix (the usual data for AMOVA) is computationally unwieldy, so we opt for a simpler (but mathematically equivalent) method of computation. The first step in the analysis is to compute the distance between male gametes; we assume that the distance between any two alleles at a single locus is the same, namely '1'. A simple coding of the pollen data, representing each allele as a separate row in a vector yields the same result. For example, if a male gamete has four loci with up to four alleles per locus, we code the four-locus gamete (eg A_2 B_3 C_1 D_4) as

$$G' = \begin{matrix} [0\,1\,0\,0, & 0\,0\,1\,0, & 1\,0\,0\,0, & 0\,0\,0\,1] \\ \text{A-locus}, & \text{B-locus}, & \text{C-locus}, & \text{D-locus} \end{matrix}$$

This is convenient, because we can expand (or contract) the subvector for a particular locus, as needed, to accommodate more or fewer alleles. We can also modify this representation to describe the 'allelic state' of ambiguous male gametes. For example, if the male gamete vector were [A_2 B_3 $C_{2/3}$ D_4], where the notation $C_{2/3}$ indicates an ambiguity between allele C_2 and C_3 (as would be needed for a C_2C_3 seedling from a C_2C_3 mother), we can define the male gamete vector as

$$G' = [0\,1\,0\,0,\quad 0\,0\,1\,0,\quad 0\,\gamma_{23}\,\gamma_{32}\,0,\quad 0\,0\,0\,1]$$

where the γ_{kl} are derived from the estimated allele frequencies for the *k*th and *l*th alleles at the C locus (Appendix A). Seed tree mean vectors, yearly mean vectors and global mean vectors are obtained in the usual fashion. Squared deviations from various mean vectors are computed in the vector product form; for example for the *k*th individual in the *i*th stratum, we have

$$\delta^2_{ik} = [G_{ik} - G_{i\cdot}]'[G_{ik} - G_{i\cdot}] \qquad (1)$$

These squared deviations are summed in typical ANOVA fashion.

Two-way AMOVA design

Our strategy extends the TwoGener analysis described by Smouse *et al* (2001), using a modification of the crossclassified AMOVA layout used in Brown *et al* (1996), to fit a model that includes seed tree effects, year effects and tree × year interaction effects. We do not know a priori whether years should be treated as crossclassificatory design variables or as nested replicates within seed trees. We anticipate some year-to-year variation, but an important question is whether 'years' have any average effects, or whether they simply introduce variation within trees, without having any particular average effects. For the crossclassified design, we use the model

$$G_{ijk} = \mu + f_i + y_j + (fy)_{ij} + \omega_{ijk} \qquad (2)$$

where f_i is the average effect of the *i*th seed tree, y_j is the average effect of the *j*th year, $(fy)_{ij}$ is the interaction effect (failure of additivity) of the *i*th tree and *j*th year and ω_{ijk} is the replication error associated with the *k*th pollen gamete from the *i*th tree in the *j*th year. Effects are assumed to be random. The crossclassified random effects analysis is shown in Table 2, where we have adjusted ambiguous pollen with allele frequency estimates from each tree–year combination separately (scheme a). Using variance component extraction procedures, we produce estimates of four variance components. Results for the other three adjustment schemes (not shown) yield the same basic result; averaging over years reduces the interaction variance of trees and years, and averaging over trees reduces the variance among trees. All four adjustment methods show that main effects for years are nil; they can be safely ignored.

Three-level nested design

Although the year-to-year main effects are not significant, nontrivial temporal variation within a tree indicates that a simpler (three-level nested) model is in order

$$G_{ijk} = \mu + f_i + y_{j(i)} + \omega_{ijk} \qquad (3)$$

where f_i is the average effect of the *i*th seed tree, $y_{\mathrm{j(i)}}$ is the average effect of the *j*th year, nested within the *i*th tree and ω_{ijk} is the replication error associated with the *k*th pollen gamete from the *i*th tree in the *j*th year. Again, all effects are assumed to be random. The form of the nested analysis is shown in Table 3.

Results from the doubly nested model (Table 3) can be contrasted with those of the crossclassified model (Table 2). To compute the sum of squares for 'years within trees' in Table 3, the sums of squares for 'years' and 'interaction' from the crossclassified analysis (Table 2) are summed. Similarly, to obtain the degrees of freedom for 'years, nested within trees' in Table 3, the degrees of freedom (3 and 42) for the 'years' and 'interaction' effects in Table 2 are summed. The 'among trees' component of variation is large (17.4% of the total variation). Year-to-year variation for a given maternal tree, while somewhat smaller, is also significant and not trivial (6.9%).

Table 2 Cross-classified analysis of molecular variation for pollen variation among 15 *Albizia julibrissin* Durazz. trees in the Athens area, each sampled in four successive years, and with 40 seedlings sampled per year

Source of variation	*d. f.*	*SS*	*MS*	*Expected mean squares*	*Variance component estimates*	$\hat{\sigma}^2$	$\%\hat{\sigma}^2_{tot}$
Among trees	14	886.4	63.3	$\sigma^2_{\mathrm{w}}+40\sigma^2_{\mathrm{fy}}+160\sigma^2_{\mathrm{f}}$	$\hat{\sigma}^2_{\mathrm{f}}$=(MS$_{\mathrm{f}}$−MS$_{\mathrm{fy}}$)/160	0.351	17.4
Among years	3	19.30	6.44	$\sigma^2_{\mathrm{w}}+40\sigma^2_{\mathrm{fy}}+600\sigma^2_{\mathrm{y}}$	$\hat{\sigma}^2_{\mathrm{y}}$=(MS$_{\mathrm{y}}$−MS$_{\mathrm{fy}}$)/600	−0.001	−0.1
Trees × years	42	301.2	7.17	$\sigma^2_{\mathrm{w}}+40\sigma^2_{\mathrm{fy}}$	$\hat{\sigma}^2_{\mathrm{fy}}$=(MS$_{\mathrm{fy}}$−MS$_{\mathrm{w}}$)/40	0.141	7.0
Error	2340	3566.2	1.52	σ^2_{w}	$\hat{\sigma}^2_{\mathrm{w}}$=MS$_{\mathrm{w}}$	1.52	75.6

Table 3 Nested analysis of molecular variation for pollen variation among 15 *Albizia julibrissin* Durazz. trees in the Athens area, each sampled in 4 successive years, and with 40 seedlings sampled per year

Source of variation	*d. f.*	*SS*	*MS*	*Expected mean squares*	*Variance component estimates*	$\hat{\sigma}^2$	$\%\hat{\sigma}^2_{tot}$
Among trees	14	886.4	63.3	$\sigma^2_{\mathrm{w}}+40\sigma^2_{\mathrm{y}}+160\sigma^2_{\mathrm{f}}$	$\hat{\sigma}^2_{\mathrm{f}}$=(MS$_{\mathrm{f}}$−MS$_{\mathrm{y}}$)/160	0.351	17.4
Years/trees	45	320.5	7.12	$\sigma^2_{\mathrm{w}}+40\sigma^2_{\mathrm{y}}$	$\hat{\sigma}^2_{\mathrm{yf}}$=(MS$_{\mathrm{y}}$−MS$_{\mathrm{w}}$)/40	0.140	6.9
Within trees	2340	3566.2	1.52	σ^2_{w}	$\hat{\sigma}^2_{\mathrm{w}}$=MS$_{\mathrm{w}}$	1.52	75.6

Pollen structure and the effective pollen pool

We can now estimate various intraclass correlation coefficients, Φ, from Table 3. We might be tempted to describe the proportion of the variance among trees as

$$\hat{\Phi}_{\mathrm{ft}}=\frac{\hat{\sigma}^2_{\mathrm{f}}}{\hat{\sigma}^2_{\mathrm{w}}+\hat{\sigma}^2_{\mathrm{f}}}$$

(Smouse *et al*, 2001; Austerlitz and Smouse, 2001), but based on our results, a single-year study has interyear variation (within a tree) confounded with intertree variation, and the estimator reflects that. A single-year study only allows us to estimate

$$\hat{\Phi}_{\mathrm{yt}}=\frac{\hat{\sigma}^2_{\mathrm{y}}+\hat{\sigma}^2_{\mathrm{f}}}{\hat{\sigma}^2_{\mathrm{w}}+\hat{\sigma}^2_{\mathrm{y}}+\hat{\sigma}^2_{\mathrm{f}}}=0.244 \qquad (4)$$

which would (incorrectly) lead to an estimate of $\hat{N}_{\mathrm{ep}}\sim(2\Phi_{\mathrm{yt}})^{-1}=2.05$. What is needed is the correlation of uniting male gametes for a single tree, relative to the total pollen draw of all 15 trees, with allowance for year-to-year variation within a tree, measured as the proportion of the total variance among trees. This requires data from at least 2 years, and is given by

$$\hat{\Phi}_{\mathrm{ft}}=\frac{\hat{\sigma}^2_{\mathrm{f}}}{\hat{\sigma}^2_{\mathrm{w}}+\hat{\sigma}^2_{\mathrm{y}}+\hat{\sigma}^2_{\mathrm{f}}}=0.174\pm0.062 \qquad (5)$$

with a standard error as calculated below. The revised 4-year estimate yields $\hat{N}_{\mathrm{ep}}=2.87$ (the 95% CI is [1.69, 9.41]). The value of $\hat{N}_{\mathrm{ep}}=2.87$ is the maximum effective number of pollen donors, since it accounts for the yearly variability assessed by our 4-year survey.

Another intraclass correlation of interest is that among years, within a tree, measured as the proportion of the variation within a tree that represents the year-to-year variation,

$$\hat{\Phi}_{\mathrm{yf}}=\frac{\hat{\sigma}^2_{\mathrm{y}}}{\hat{\sigma}^2_{\mathrm{w}}+\hat{\sigma}^2_{\mathrm{y}}}=0.084\pm0.041 \qquad (6)$$

Thus, about 8.4% of the variation within a tree is because of differences among the successful pollen donors from year to year. The three coefficients are related in the usual '*F*-statistic' fashion,

$$(1-\Phi_{\mathrm{yt}})=(1-\Phi_{\mathrm{yf}})(1-\Phi_{\mathrm{ft}}). \qquad (7)$$

A single-year study produces Φ_{yt}, but what is needed is Φ_{ft}; for that, we need 2 years (and probably more) of data on the same maternal trees.

Discussion

In this study, there is no net year-to-year variation in pollen allele frequencies, averaged over seed trees. This is not too surprising since mimosa (unlike many other tree species) flowers profusely every year. There is, however, significant yearly variation in the pollen pool of individual trees. For mimosa, the among-year variance component is ~40% of the intertree variance; for other species, it could be larger, especially for species that experience mast flowering. In general, to obtain accurate estimates of N_{ep}, we must account for the interyear variance of individual trees, which means we should:

- sample each individual over at least 2 (preferably more) years, so that meaningful estimates of year-to-year variation within a plant can be extracted;
- use yearly average (global) pollen frequencies to adjust ambiguous gametic vectors for all plants; and
- build the year-to-year variation explicitly into per-generation estimates of Φ_{ft} and N_{ep}.

In the present case, a single year's data would have led us to $\Phi_{\mathrm{yt}}\sim0.244$ and $N_{\mathrm{ep}}\sim2.05$, whereas allowance for year-to-year variance in the pollen profile yields a smaller estimate of $\Phi_{\mathrm{ft}}\sim0.174\pm0.062$ and a larger estimate of $N_{\mathrm{ep}}\sim2.87$ (the 95% CI is [1.69, 9.41]). Thus, the profile of pollen donors increases with multiyear sampling, and such sampling has a meaningful influence on our estimate of the effective number of pollen donors. In simpler terms, our data suggest that, on average, a given tree samples a different array of pollen donors each year, but that most of the successful pollen fertilizing an individual comes from the same (very few) pollen donors (or donors carrying the same alleles) year after year. Minor contributors vary, but major contributors do not. Although we had no a priori expectations, we could easily have imagined that a predominantly different (nonoverlapping) group of pollen donors would be sampled by individual seed trees each year. The data strongly suggest otherwise. The observed pattern could be attributed to several characteristics of mimosa. First, mimosa appears to be self-incompatible (Godt and Hamrick, 1997). Second, these 15 trees occurred in small clusters of trees (2–10 individuals) that may be related because of multiseeded fruit dispersal (as opposed to individual seed dispersal). The occurrence of full-sibs within such fruits would increase the probability that individuals would share incompatibility alleles. Also, many clusters likely develop from the recruitment of progeny from one or a few colonists. It is certainly feasible that matings between members of a cluster could be limited by the number and distribution of self-incompatibility alleles present in the cluster. Formal paternity analysis procedures (currently under way) will test this speculation. In addition, the low interyear variation in flower production should dampen the interyear variance in the pollen-donor pool. Finally, since each of the 15 trees included in this analysis was a member of a cluster of trees, a member of a larger cluster might be expected to have a different N_{ep} value. One might expect, for example, that isolated trees would receive pollen from a wider array of pollen donors and that N_{ep} for such individuals might be higher than obtained in this study.

Optimal allocation of sampling effort

Our TwoGener analysis demonstrates that year-to-year variation in pollen pools for a particular tree is substantial, indicating that it is important to obtain multiple-year samples from perennial plants. Since the sampling and assay effort for such analyses is already large and costly, it would be useful to keep the total sampling effort to a minimum. Thus, the question of optimal sample allocation arises. How many seed plants should be sampled over how many reproductive seasons each and how many seedlings should be sampled per season? We obviously need to sample at least a few seedlings in each of at least a few reproductive years to provide a basis for estimation (ie, n seedlings per year for each plant, in each of m years). Since we want to estimate Φ_{yf}, what is the optimal allocation of n and m, given a fixed number of seedlings sampled per plant ($K = mn$)? Also, since we are interested in Φ_{ft}, what is the optimal allocation of total sampling effort within- and among-plants? In other words, for a fixed total sample size, $N = KJ$, how should we choose the number of plants (J) and the number of seedlings per plant (K)?

Optimal allocation will obviously depend on the relative sizes of three variance components (σ_{f}^2, σ_{y}^2 and σ_{w}^2). We have only the current study to use as guidance, but given a set of criteria to be estimated, standard strategies are available (eg Searle *et al*, 1992). Given our results, it would probably be reasonable to assume that $9\sigma_{\mathrm{y}}^2 \sim 4\sigma_{\mathrm{f}}^2 \sim \sigma_{\mathrm{w}}^2$, but for other situations or species, we can generalize the argument by setting $\sigma_{\mathrm{y}}^2 = r\sigma_{\mathrm{w}}^2$ and $\sigma_{\mathrm{f}}^2 = t\sigma_{\mathrm{w}}^2$. First, we must define the criteria to be optimized. The two parameters we most wish to estimate from this three-level nested AMOVA, and their translations into r and t, are:

$$\Phi_{\mathrm{yt}} = \frac{\sigma_{\mathrm{y}}^2}{\sigma_{\mathrm{w}}^2 + \sigma_{\mathrm{y}}^2} = \frac{r}{1+r} \quad \text{and}$$

$$\Phi_{\mathrm{ft}} = \frac{\sigma_{\mathrm{f}}^2}{\sigma_{\mathrm{w}}^2 + \sigma_{\mathrm{y}}^2 + \sigma_{\mathrm{f}}^2} = \frac{t}{1+r+t} \qquad (8)$$

Variances of the estimates of such criteria can be derived (Searle *et al*, 1992), and they are (lengthy algebra not reproduced here):

$$\mathrm{Var}(\hat{\Phi}_{\mathrm{yf}}) = \frac{2}{Jn^2}\Phi_{\mathrm{yf}}^2\{A+B\} \quad \text{and}$$

$$\mathrm{Var}(\hat{\Phi}_{\mathrm{ft}}) = \frac{2}{K^2}\Phi_{\mathrm{ft}}^2\{C+D+E\} \qquad (9)$$

with A, B, C, D and E defined as

$$A = \frac{(1+nr)^2}{(m-1)r^2}(1-\Phi_{\mathrm{yf}})^2,$$

$$B = \frac{1}{m(n-1)}\left(\frac{1}{r} + \frac{(n-1)}{(1+r)}\right)^2$$

$$C = \frac{(1+nr+Kt)^2}{(J-1)t^2}(1-\Phi_{\mathrm{ft}})^2,$$

$$D = \frac{(1+nr)^2}{J(m-1)}\left[\frac{1}{t} + \frac{(m-1)}{(1+r+t)}\right]^2 \qquad (10)$$

$$E = \frac{m(n-1)}{J(1+r+t)^2}$$

These variances assume that we know the parametric r and t values, which were chosen to generalize the argument. To obtain unbiased estimators of the coefficients we have specifically estimated for mimosa, we add 2 to the degrees of freedom in the mean square estimators. Having done that, estimated confidence intervals are those reported above.

The object of the planning exercise is to allocate future sampling efforts in such a fashion as to obtain estimates of Φ_{yf} and Φ_{ft} with minimal standard deviations, relative to their parametric mean values. A bit of manipulation yields the pair of criteria to minimize

$$\mathrm{CV}(\Phi_{\mathrm{yf}}) = \frac{\sqrt{2(A+B)/J}}{n} \quad \text{and}$$

$$\mathrm{CV}(\Phi_{\mathrm{ft}}) = \frac{\sqrt{2(C+D+E)}}{mn} \qquad (11)$$

relative to the choices of J, m and n, for fixed total sample size, $N = JK = J(mn)$. The recommendations for minimizing $\mathrm{CV}(\Phi_{\mathrm{yf}})$ work at cross purposes to those for minimizing $\mathrm{CV}(\Phi_{\mathrm{ft}})$, but only mildly so. For a fixed ratio of m:n, the best way to optimize $\mathrm{CV}(\Phi_{\mathrm{yf}})$, for a fixed total sample size, N, is to increase the number of seedlings per maternal plant, at the expense of the number of plants, J. For that same m:n ratio and total N, one should increase the number of plants to minimize $\mathrm{CV}(\Phi_{\mathrm{ft}})$, reducing the number of seedlings per plant, K, to a minimum (Figure 1a). On the other hand, $\mathrm{CV}(\Phi_{\mathrm{yf}})$ is relatively insensitive to the J:K ratio, compared with $\mathrm{CV}(\Phi_{\mathrm{ft}})$, which declines rapidly, as J:K increases, so there is some flexibility.

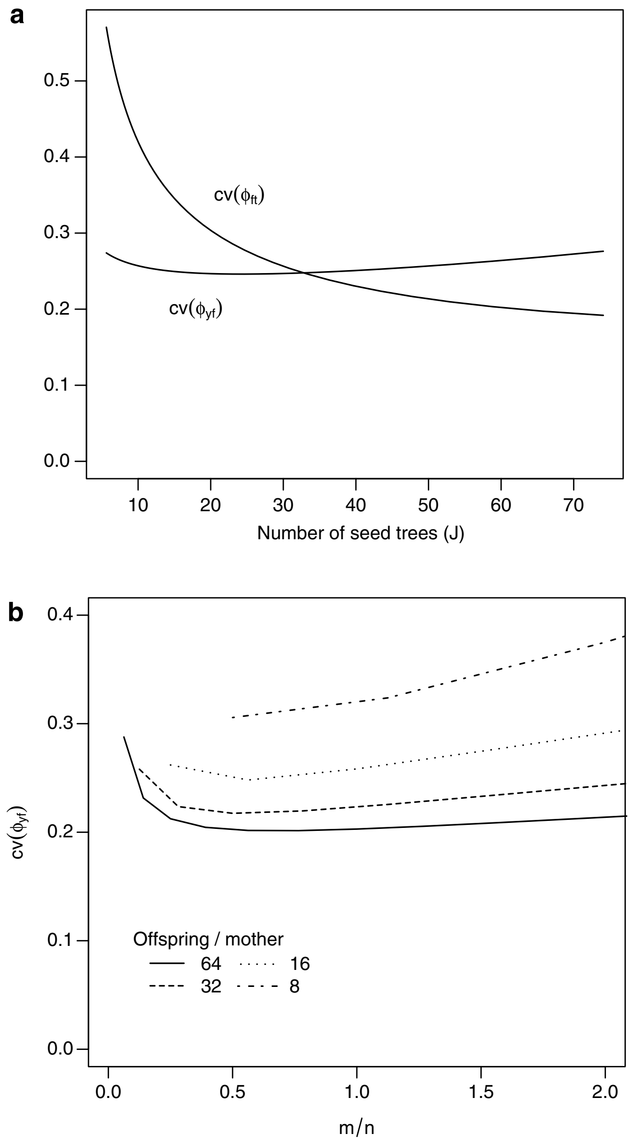

Figure 1 Effect of reallocating sampling effort on CV(Φ_{yf}) and CV(Φ_{ft}), keeping the total sample size ($N=2400$) constant: (a) Increasing the number of seed trees (*J*) decreases CV(Φ_{ft}), while increasing CV(Φ_{yf}) slightly. (b) Increasing the number of sample years at the expense of replication within years, while keeping the total number of trees constant shows a minimum CV(Φ_{yf}) near $m/n=1/2$ for studies similar to ours.

For any fixed value of *K* and *J*, CV(Φ_{yf}) is sensitive to the *m*:*n* ratio (Figure 1b), for any given value of *K* (hence, *J*). There are clearly tradeoffs to be considered. Our Athens mimosa design of $m=4$ years, $n=40$ seedlings per year and $J=15$ seed trees, yields CV(Φ_{yf}) ~ 0.250 and CV(Φ_{ft}) ~ 0.346, adequate for the testing purposes at hand, but probably larger than ideal. Had we initially sampled more seed trees (*J*), with fewer progeny (*K*) within trees to compensate, we would have achieved better precision. In retrospect, a design with $m=4$, $n=8$, $J=75$, for example, would have yielded CV(Φ_{yf}) ~ 0.217 and CV(Φ_{ft}) ~ 0.167, with no increase in total sample size. Optimal allocation strategies are a bit sensitive to changes in the unknown *r* and *t* values, of course, but in deference to our data, we have used $r \sim 0.09$ and $t \sim 0.25$ for this evaluation. With those values of *r* and *t*, we discover that a design with ($m=4$, $n=8$, $J=75$) would have provided a 95% CI on N_{ep}, on the order of [2.15, 4.30], a substantial improvement on our current precision. It seems clear that preliminary data on the relative sizes of the variance components will be very much worthwhile for planning purposes.

Ambiguous pollen designations

Male gametophytic (pollen) genotypes were determined by subtracting the contribution of the maternal genotype. If the offspring and seed tree have the same heterozygous genotype, their contributions are ambiguous. The gametic contributions of ambiguous pollen can be deduced (on average) from the unambiguous allele frequencies, but which set of average allele frequencies to use is a matter of judgment. We had large intraplant sample sizes here and we elected to use the averages for each tree-year combination. The sparser within-tree replication that will accompany alternative designs may make it necessary to use yearly, seed tree, or even global averages. We have evaluated those schemes here as well (not shown), and the results were similar for alternative averaging schemes. Broad averaging inevitably favors the null hypothesis, but sample size considerations argue in favor of that practice (Smouse *et al*, 2001). With fewer progeny per tree in future studies, the matter needs further attention.

We have addressed the question of whether sampling across multiple reproductive events provides insights into the lifetime pollen-donor pools of mimosa. Our results demonstrate that adding reproductive episodes to the study increases estimates of the effective number of pollen parents, much the same way sampling more individuals in a population increases the number of alleles observed. Furthermore, our results have demonstrated that increasing the number of reproductive events studied had a meaningful impact on estimates of the *effective* number of pollen donors. In this multiyear analysis, the estimated effective number of pollen parents increased by 40% over that obtained by sampling a single reproductive event. For tree species with more heterogeneous interyear flowering, we would expect the number of pollen parents to increase even more with multiyear samples. At the very least, the substantial increase seen for mimosa indicates that multiyear analyses are probably necessary for accurate estimates of the effective number of pollen donors sampled during the lifetime of perennial plants.

Acknowledgements

We to thank VL Sork, RD Westfall, RJ Dyer and F Austerlitz for helpful commentary on the manuscript. Susan Sherman-Broyles provided valuable technical assistance for this study. JLH and MJWG were supported by USDA Grant # H2000–00882. AJI and PES were supported by the New Jersey Agriculture Experiment Station, USDA/NJAES-32106, by McIntire Stennis-17309, and by NSF-DEB-0089238.

References

Austerlitz F, Smouse PE (2001). Two-generation analysis of pollen flow across a landscape. II. Relation between Φ_{ft}, pollen dispersal and inter-female distance. *Genetics* **157**: 851–857.

Brown BL, Epifanio JM, Smouse PE, Kobak CJ (1996). Temporal stability of mtDNA haplotype frequencies in American shad stocks: to pool or not to pool across years? *Can J Fish Aquat Sci* **53**: 2274–2283.

Broyles SB, Wyatt R (1990). Paternity analysis in a natural population of *Asclepias exaltata*: multiple paternity, functional gender and the 'pollen-donation hypothesis'. *Evolution* **44**: 1454–1468.

Chakravarti A, Li CC (1983). The effect of linkage on paternity calculations. In: Walker RH (ed) *Inclusion Probabilities in Parentage Testing*, American Association Blood Banks: Arlington, VA, pp 411–422.

Devlin B, Ellstrand NC (1990). Male and female fertility in wild radish, a hermaphrodite. *Am Nat* **136**: 87–107.

Elias TS (1987). *Trees of North America*. Gramercy Publishing: New York.

Ellstrand NC, Marshall DL (1985). Interpopulation gene flow by pollen in wild radish, *Raphanus sativus*. *Am Nat* **126**: 606–616.

Excoffier L, Smouse PE, Quattro JM (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction sites. *Genetics* **131**: 479–491.

Godt MJW, Hamrick JL (1993). Patterns and levels of pollen-mediated gene flow in *Lathyrus latifolius*. *Evolution* **47**: 98–110.

Godt MJW, Hamrick JL (1997). Estimation of mating system parameters of *Albizia julibrissin* (Fabaceae). *For Genet* **4**: 217–221.

Hamrick JL, Murawski DA (1990). The breeding structure of tropical tree populations. *Pl Spec Biol* **5**: 157–165.

Jamieson A (1994). The effectiveness of using co-dominant polymorphic allelic series for (1) checking pedigrees and (2) distinguishing full-sib pair members. *Anim Genet* **25**: 37–44.

Meagher TR (1986). Analysis of paternity within a natural population of *Chamaelirium luteum*. I. Identification of most-likely male parents. *Am Nat* **128**: 199–215.

Schemske DW, Husband BC, Ruckelhaus MH, Goodwille C, Parker IM, Bishop JG (1994). Evaluating approaches to the conservation of rare and endangered plants. *Ecology* **75**: 584–606.

Schnabel A, Hamrick JL (1995). Understanding the population genetic structure of *Gleditsia triacanthos* L.: the scale and pattern of pollen gene flow. *Evolution* **49**: 921–931.

Schoen DJ, Stewart SC (1986). Variation in male reproductive investment and male reproductive success in white spruce. *Evolution* **36**: 352–360.

Searle SR, Casella G, McCulloch CE (1992). *Variance Components*. Wiley & Sons: New York.

Smouse PE, Dyer RJ, Westfall RD, Sork VL (2001). Two-generation analysis of pollen flow across a landscape. I. Male gamete heterogeneity among females. *Evolution* **55**: 260–271.

Sork VL, Campbell D, Dyer R, Fernandez J, Nason J, Petit R, Smouse P, Steinberg E (1998). Proceedings from a workshop on gene flow in fragmented, managed, and continuous populations. National Center for Ecological Analysis and Synthesis, Santa Barbara, CA. Res. Pap. 3. http://www.nceas.ucsb.edu/nceas-web/projects/2057/nceas-paper3/.

Wright S (1951). The genetical structure of populations. *Ann Eugen* **15**: 323–354.

Appendix A

Reconstruction of Pollen Allele Frequencies when the Maternal Genotype is Heterozygous

If the mother and offspring share the same heterozygous genotype, say C_2C_3, it is impossible to determine which allele came from the pollen and which came from the ovule. When we first calculate pollen allele frequencies, we discard these ambiguous alleles, but this reduces the sample size and biases the allele frequencies. For example, consider the C-locus, with four alleles {C_1, C_2, C_3 and C_4}, with frequencies p_1, p_2, p_3 and p_4, respectively. Assume the maternal genotype is C_2C_3. The frequencies of the eight combinations of ovule and pollen are:

	Paternal Gametes				
Maternal Gametes	C_1	C_2	C_3	C_4	
C_2	$p_1/2$	$p_2/2$	**$p_3/2$**	$p_4/2$	½
C_3	$p_1/2$	**$p_2/2$**	$p_3/2$	$p_4/2$	½
	p_1	p_2	p_3	p_4	

Ignoring the ambiguous offspring (bold) produces estimates of pollen allele frequency P_i', which can be used to fill in the holes left by the ambiguities. Since the ambiguous pollen is either C_2 or C_3, ambiguous pollen genotypes can be assigned a proportional pollen frequency vector,

$$P_i' = [0 \quad \hat{\gamma}_{23} \quad \hat{\gamma}_{32} \quad 0] \tag{A.1}$$

where

$$\hat{\gamma}_{23} = \frac{\hat{p}_2}{\hat{p}_2 + \hat{p}_3} \quad \text{and} \quad \hat{\gamma}_{32} = \frac{\hat{p}_3}{\hat{p}_2 + \hat{p}_3} \tag{A.2}$$

If we use a P-vector of this type for any ambiguous pollen, allele frequencies for the population will be unbiased and we recover the information provided by the ambiguous alleles. This procedure yields the maximum likelihood estimates of the γ_{ij}. When determining the sample size, we then have a full set of data. However, since some of the allele frequencies are estimated from a smaller sample, the variance will be elevated.