Theor Appl Genet (1994) 88:593–596


S. C. Stewart

# Simultaneous estimation of pollen contamination and pollen fertilities of individual trees in conifer seed orchards using multilocus genetic data

Received: 12 September 1993 / Accepted: 12 October 1993

**Abstract** Seed orchards of forest trees are established to produce genetically-improved seed for reforestation. Genetic efficiency requires seed orchards to be (1) reproductively isolated form surrounding trees, (2) that there be similar fertilities among all orchard trees, and (3) minimum inbreeding. Each aspect of seed orchard reproduction can be simultaneously estimated using the observed frequency of each multilocus gametic type contributed through fertilizing pollen and the expected multilocus gametic segregation frequencies of orchard tree and of the contaminating population. These genetic estimates are directly relevant to the genetics of the tree breeding program. The flexibility of sampling seed – the basic data for these techniques – allows great scope for hypothesis testing, including tests of the accuracy of predictions of biophysical models of pollen movement. A simple example and a white spruce seed orchard case study are presented to illustrate the estimation technique and to investigate its sensitivity.

**Key words** Male fertility · Genetic efficiency · Pollen contamination · Conifer seed orchard · Inbreeding

## Introduction

Seed orchards of forest trees are established to produce genetically-improved seed for reforestation. Genetic efficiency and avoidance of inbreeding are often the two most important areas of orchard quality control. Genetic efficiency requires seed orchards to be reproductively isolated from surrounding trees and with close to equal gametic contributions to the seed crop from each individual in the orchard. Without reproductive isolation genetic gain is compromised to the extent that non-selected trees contribute gametes to the orchard seed crop. Unequal fertilities among selected trees within the orchard, especially unequal pollen fertilities, also generally lower the expected genetic gain of the breeding program. This is because orchard trees with desirable characteristics, but low pollen fertility, fail to contribute gametes coding for their traits to the orchard seed crop. Inbreeding can also severely reduce seed quality in most conifer species.

Each of these aspects of reproduction – pollen contamination, pollen fertilities among selected parents, and inbreeding – can be estimated using genetic data. Perhaps the most significant advantage of genetic estimation techniques is the direct relationship between the estimation of quantitative genetic parameters which form the basis of the orchard's selective breeding program and the genetic estimates of contamination, pollen fertilities, and inbreeding. For example, pollen which enters the orchard (and even orchard seed cones), but which fails to fertilize an ovule (and which, therefore, is of no genetic consequence), will not appear in the sample, and will not inflate the estimate of contamination. Thus, genetic estimation techniques are often the most accurate, precise, and appropriate technique for tree breeding and orchard management decisions.

Genetic-estimation techniques identify and exploit genetic differences among the various trees contributing gametes to the orchard seed crop. Conifer seed are especially convenient because the haploid megagametophyte in each seed is genetically identical to the maternal parent's gametic contribution to the embryo. With this information the multilocus diploid genotype of the maternal parent can be reliably inferred from a small sample of seed. The multilocus genotype of the fertilizing pollen gamete can also be determined. This is possible because the pollen genotype is comprised of alleles in the diploid seed embryo (which includes both the gametic contribution from the fertilizing pollen grain *and* the gametic contribution of the mother) which are not in the haploid seed megagametophyte (which includes the gametic contribution from the maternal

Communicated by C. Smith

S. C. Stewart
Department of Botany, University of Guelph, Guelph, Ontario N1G 2W1, Canada

parent only). For these reasons the genotypes of seed generally form the basic data for genetic estimation techniques.

There is also remarkable flexibility in sampling seed, which offers scope for hypothesis testing. Consider, for example, the important case of testing the accuracy of biophysical models of pollen movement. Such models (e.g., Di-Giovanni and Beckett 1990; Di-Giovanni and Kevan 1991 and references therein) claim to predict the distribution and abundance of pollen within orchards (and forests). The accuracy of the predictions can be tested by comparing contamination estimates for separate areas of an orchard to the predicted pollen load for each location. This requires only independent samples of seed from each location. Reliable estimates with 95% confidence intervals that are less than 3% of the estimate typically require a sample of 1 000 seed. However, samples of as few as 100 seed often provide useful (but less precise) estimates. Other testing examples include, estimates of within orchard pollen dispersal distances (Schoen and Stewart 1986, 1987; Yazdani et al. 1989), estimates of pollen fertility in controlled crosses using specific pollen mixtures, and management practices designed to reduce contamination and self-fertilization rates, such as supplemental-mass-pollination and the use of water sprays to cool orchard cones, slow development, and delay orchard reproduction to after pollen release of non-orchard pollen sources (El-Kassaby and Ritland 1986).

A useful example of how estimates can be calculated is illustrated by the case where the pollen of each tree is genetically distinct and unique. In this case determining the genotype of the pollen unambiguously identifies whether the pollen was produced by a non-orchard tree (contamination), an orchard tree other than the seed parent (pollen fertility), or by the seed parent (self-fertilization). In this case, estimation is a straightforward count of fertilization events. Unfortunately, such unique markers are rare. However, early approaches did focus on such unique genetic differences. For example, Friedman and Adams (1981) identified allozyme alleles unique to the surrounding natural stands in their loblolly pine orchard and assayed orchard progeny for these rare alleles. Squillace and Long (1981) identified recombinant genotypes in the orchard progeny which could not have been produced from legitimate orchard crosses. Smith and Adams (1983) used a similar approach. The rarity of unique markers greatly limits the precision and applicability of these early estimation techniques.

Multilocus markers have been used to assign paternity of seed sampled from individual mother trees (Neale 1984; Hamrick and Schnabel 1985; Meagher 1986; Devlin et al. 1988). Paternity analysis depends on identifying the pollen parent by genetic exclusion. Unfortunately, complete genetic exclusion is rarely possible and without complete exclusion paternity estimates are expected to be biased (Brown et al. 1985; Brown 1989).

In 1986, estimation techniques were introduced that model the multilocus probability structure of the entire sample of seed (or pollen) genotypes to estimate mating system parameters. El-Kassaby and Ritland (1986) described a method to estimate contamination based upon differences of single-locus allele frequency between orchard and the non-orchard pollen sources. This was an improvement over earlier techniques because estimates incorporated information provided by allelic differences unique to each pollen source and also quantitative allelic differences. This increased utilization of the genetic information greatly improves the precision of the estimates. Schoen and Stewart (1986) presented a probability model that estimates individual male fertilities that avoids the bias expected with paternity analysis, but did not account for pollen contamination. Adams and co-workers developed a probability model to estimate the population average selfing rate, contamination rate, and a distance parameter that summarizes the role of distance in determining pollen fertility (Adams and Birkes 1989, 1991; Adams et al. 1992). Selfing and contamination are important aspects of seed orchard management. However, the genetic efficiency of tree breeding in seed orchards also depends on the mating success of individual orchard trees, which the Adams-Birks model does not estimate. In this paper I present a probability model technique based on multilocus gamete frequency differences among pollen sources that simultaneously estimates pollen contamination, pollen fertilities, and selfing rates of orchard trees.

## The estimation model

Let $i = (i_1, i_2, \ldots, i_n)$ represent the $n$-locus gamete, where $i_a$ represents the allele at locus $a$. Next, let $p_i$ be the frequency of gamete $i$ contributed as pollen to seed. Each pollen gamete frequency is the sum of the contribution from individuals in the orchard and the contribution from (non-orchard) individuals in the contaminating population. The contribution from individuals in the orchard can be denoted as $\sum R(i,jk) f_{jk}$, where $R(i,jk)$ is the known probability that individuals comprised of gametes $j$ and $k$ produce gamete $i$, $f_{jk}$ is the proportion of seed sired by orchard individuals comprised of gametes $j$ and $k$, and summation is over all gametic types $j$ and $k$. The contribution from individuals in the contaminating population can be denoted as $p'_i c$, where $p'_i$ is the known frequency of gamete $i$ produced by individuals of the contaminating population and $c$ is the proportion of seed sired by those non-orchard individuals. The frequency of gamete $i$ contributed as pollen to seed from both sources is, therefore,

$$p_i = (\sum R(i,jk) f_{jk}) + p'_i c, \tag{1}$$

for all gamete types $i, j, k$. In matrix notation, the set of equations can be represented as

$$\mathbf{p} = (\mathbf{R}|\mathbf{p}')(\mathbf{f}^{\mathrm{T}}|\mathbf{c}^{\mathrm{T}})^{\mathrm{T}} \tag{2}$$

where $\mathbf{p}$ is a column vector containing the frequency of each $n$-locus gamete of type $i$ contributed as pollen to seed, $(\mathbf{R}|\mathbf{p}')$ is a partitioned matrix comprised of $\mathbf{R}$ (a matrix of gametic segregation frequencies where rows are indexed by $n$-locus gamete of type $i$ and columns by $n$-locus diploid genotypic class $jk$) and $\mathbf{p}'$ (a column vector containing

the known frequency of each $n$-locus gamete of type $i$ produced by trees in the contaminating population), $\mathbf{f}$ is a column vector containing the male fertility of each $n$-locus diploid genotypic class $jk$, $\mathbf{c}$ is a vector of contamination rates (generally constrained to be the same for all gamete types), and $^{T}$ denotes transposition. Equation (2) always has a solution,

$$(\mathbf{f}^{T}|\mathbf{c}^{T})^{T} = \mathbf{G}\mathbf{p} \tag{3}$$

where $\mathbf{G}$ is the generalized inverse of $(\mathbf{R}|\mathbf{p}')$ (Searle 1971). Note that, if the elements of $\mathbf{G}$ [and equivalently, $(\mathbf{R}|\mathbf{p}')$] are real constants, then

$$E(\mathbf{f}^{T}|\mathbf{c}^{T})^{T} = \mathbf{G}E(\mathbf{p}), \tag{4}$$

where $E(\ )$ denotes expectation (Elant-Johnson 1971). Therefore, solving equation (4) provides estimates of orchard individual male fertilities and the rate of contamination directly from the observed frequencies of the $n$-locus gamete types contributed as pollen to some set of seeds whenever gametic segregation frequencies are known.

In general, the solution of equation (4) need not be unique. However, when $(\mathbf{R}|\mathbf{p}')$ (and equivalently, $\mathbf{G}$) is a matrix of rank $r$ there exists a set of linearly-independent linear functions of the male-success parameters for which an estimator can be found that is invariant to whatever solution of equation (4) is used; such functions are called estimable [for a discussion see, e.g., Searle (1971)]. One such set of linearly-independent estimable functions can be obtained from the row echelon form of $(\mathbf{R}|\mathbf{p}')$, where each row contains the set of coefficients of one estimable function. Rows which have a single nonzero element indicate that the proportion of seed sired by the corresponding genotypic class can be estimated independently of all other genotypic class. Rows with more than one non-zero element indicate that some combinations of male fertilities of more than one genotypic class (and, perhaps, contamination) can be estimated. A simple example illustrating the estimation technique is given in the Appendix.

The variance of each estimable function of pollen fertilities and contamination can be calculated directly from the variances and convariances of the gamete frequencies, using the standard formula for the variance of a linear function of random variables (Elant-Johnson 1971; Searle 1971). If fertilization events are uncorrelated, gamete frequencies are expected to be multinomially distributed.

Several features of the estimation procedure should be summarized. First, the expected gametic segregation frequencies must be known and incorporated into the estimation model. Second, the procedure requires estimates of male gamete frequencies from within orchard seed (estimated from a sample of seed taken from the orchard) and estimates of male gamete frequencies produced in the contaminating population (estimated from a sample of seed taken from the contaminating population). Pollen and ovule gametes of the seed must, therefore, be distinguishable. Third, mating events are assumed to be independent. Fourth, the resolution of the procedure increases with additional polymorphic loci. It is, therfore, possible to estimate the pollen fertilities of each individual in a population and contamination, if sufficient genetic variation exists, if it is assayed, and if the multilocus genotype of all "legitimate" pollen parents are known. Otherwise, the model provides estimates for each multilocus genotypic class and contamination. Fifth, the set of pollen fertilities and contamination can be estimated separately for any set of seed of interest. This offers the flexibility needed to investigate very specific hypotheses. For example, the selfing rate of a single genetic individual can be estimated from seed collected from that individual alone. In this case, the pollen fertility of the seed parent is the estimate of the selfing rate of that tree. Finally, the estimation procedure obviates the problems associated with attributing ambiguous progeny fractionally to all possible pollen parents in proportion to their likelihoods (Brown et al. 1985; Adams and Birks 1989, 1991; Adams et al. 1992), because all genotypic classes with nonzero pollen fertility estimates have been observed to contribute male gametes.

## A seed-orchard case study

Pollen fertility variation, selfing rate variation, and contamination were estimated for a clonal seed orchard of white spruce located in Glencairn, Ontario, Canada. Pollen fertility and selfing-rate variation was previously estimated assuming no pollen contamination (Schoen and Stewart 1986, 1987). This example offers the opportunity to relax the assumption of no contamination, to investigate the senesitivity of fertility estimates to simultaneously estimating contamination and male fertilities, and to detect the occurrence of long-range pollen dispersal (because there are very few non-orchard white spruce trees within 1 km of the orchard).

Description of the orchard structure, sampling methods, electrophoretic procedures, and multilocus frequencies of orchard individuals and of sampled seed appear in Schoen and Stewart (1986, 1987). The multilocus gamete frequencies of pollen produced by the contaminating population was estimated from the orchard allele frequencies assuming linkage equilibria (Epperson and Allard 1987). From these data, the rate of contamination was estimated to be $0.011 \pm 0.006$. This estimate provides little evidence that long-range pollen contamination occurred. Male fertilities of 18 orchard clones were estimated and were highly unequal among these clones: the estimated proportion of seed sired by these clones averaged 0.036 and ranged from 0.001 to 0.346. The probability that this variation was due to chance was less than 0.0001 according to chi-square testing the null hypothesis of equal male fertilities. Six complicated functions of male mating success could also be estimated but are difficult to interpret. The structure of $(\mathbf{R}|\mathbf{p}')$ determines which male fertilities can be estimated and the rank of $(\mathbf{R}|\mathbf{p}')$ determines the number of estimable parameters. In this case, $(\mathbf{R}|\mathbf{p}')$ provides enough degrees of freedom to estimate the contamination rate, 18 clonal fertities, and six estimable functions of male fertilities. None of the fertility estimates, selfing rates, or their standard errors, differed from estimates assuming no contamination by more than 3%. In conclusion, no evidence of long-range pollen contamination in this orchard was detected and estimates of male fertilities and selfing remained stable when contamination, male fertilities, and selfing rates were estimated simultaneously.

**Acknowledegements** A grant from the Natural Sciences and Engineering Research Council of Canada supported this research.

## Appendix

Let the alleles of one locus be labeled $A_1, A_2, \ldots$, and the alleles at another locus be labeled $B_1, B_2, \ldots$, Consider a seed orchard comprised of only two individuals. One individual has a two-locus genotype of $A_1B_1//A_1B_2$. The other has a two-locus genotype of $A_1B_1//A_2B_1$. Suppose the surrounding, "contaminating" population

is known to produce the four different two-locus gamete types ($A_1B_1, A_1B_2, A_2B_1, A_2B_2$) with frequencies of 3/8, 1/8, 1/8, 3/8, respectively. Thus,

$$(\mathbf{R}|\mathbf{p}') = \begin{array}{c} \begin{array}{ccc} A_1B_1//A_1B_2 & A_1B_1//A_2B_1 & \textit{"contaminating" population} \end{array} \\ \begin{pmatrix} 1/2 & 1/2 & 3/8 \\ 1/2 & 0 & 1/8 \\ 0 & 1/2 & 1/8 \\ 0 & 0 & 3/8 \end{pmatrix} \end{array} \begin{array}{l} \text{gamete } A_1B_1 \\ \text{gamete } A_1B_2 \\ \text{gamete } A_2B_1 \\ \text{gamete } A_2B_2 \end{array}$$

where the first column of **R** is the set of expected two-locus gamete frequencies for $A_1B_1//A_1B_2$, the second column for $A_1B_1//A_2B_1$, and the third column is **p'**, the set of frequencies of the contaminating population. Finally suppose the frequencies of the four two-locus gametes observed to be the male gamete among seed are 0.475, 0.225, 0.225, and 0.075, respectively.

The estimation equation, $\mathbf{p} = (\mathbf{R}|\mathbf{p}')(\mathbf{f}^T|\mathbf{c}^T)^T$, is, therefore

$$\begin{pmatrix} p_{A_1B_1} \\ p_{A_1B_2} \\ p_{A_2B_1} \\ p_{A_2B_2} \end{pmatrix} = \begin{pmatrix} 0.475 \\ 0.225 \\ 0.225 \\ 0.075 \end{pmatrix} = \begin{pmatrix} 1/2 & 1/2 & 3/8 \\ 1/2 & 0 & 1/8 \\ 0 & 1/2 & 1/8 \\ 0 & 0 & 3/8 \end{pmatrix} \begin{pmatrix} f_{A_1B_1//A_1B_2} \\ f_{A_1B_1//A_2B_1} \\ c \end{pmatrix}$$

It can be shown that one solution, $(\mathbf{f}^T|\mathbf{c}^T)^T = \mathbf{Gp}$, is

$$\begin{pmatrix} f_{A_1B_1//A_1B_2} \\ f_{A_1B_1//A_2B_1} \\ c \end{pmatrix} = \begin{pmatrix} 0.4 \\ 0.4 \\ 0.2 \end{pmatrix} \begin{pmatrix} 0.571 & 1.429 & -0.571 & -0.857 \\ 0.571 & 0.571 & 1.429 & -0.857 \\ 0.286 & 0.286 & -0.286 & 2.571 \end{pmatrix} \begin{pmatrix} 0.475 \\ 0.225 \\ 0.225 \\ 0.075 \end{pmatrix}$$

where $f_{A_1B_1//A_1B_2}$ and $f_{A_1B_1//A_2B_1}$ are the proportion of seed sired by the two orchard individuals, $c$ is the proportion of seed sired by individuals from the contaminating population, and **G** is the generalized inverse $(\mathbf{R}|\mathbf{p}')$. In conclusion, each of the individuals in the seed orchard is estimated to have fertilized 40% of seed – with equal male fertility – and 20% of the seed is estimated to have been fertilized by the surrounding population.

## References


Adams WT, Birkes DS (1989) Mating patterns in seed orchards. Proc 20th South Forest Tree Improv Conf. Charleston, South Carolina, pp 75–86

Adams WT, Birkes DS (1991) Estimating mating patterns in forest tree populations. In: Fineschi S, Malvolti M, Cannata F, Hattemer HH (eds) Biochemical markers in population genetics of forest trees. SPB Academic Publishing, The Hague, pp 157–172

Adams WT, Birkes DS, Erickson VJ (1992) Using genetic markers to measure gene flow and pollen dispersal in forest tree seed orchards. In: Wyatt R (ed) Ecology and evolution of plant reproduction. Elsevier, New York, pp 37–61

Brown AHD (1989) Genetic characterization of plant mating systems. In: Brown AHD, Clegg MT, Kahler AL, Weir BS (eds) Plant population genetics, breeding, and genetic resources. Sinauer Associates Inc, Sunderland, pp 145–162

Brown AHD, Barrett SCH, Morgan GF (1985) Mating system estimation in forest trees: models, methods and meanings. In: Gregorius HR (ed) Population genetics in forestry. Springer, Berlin Heidelberg New York, pp 32–49

Devlin B, Roeder K, Ellstrand NC (1988) Fractional paternity assignment: theoretical development and comparison to other methods. Theor Appl Genet 76:369–380

Di-Giovanni F, Beckett PM (1990) On the mathematical modelling of pollen dispersal and deposition. J Appl Meteor 29:1352–1357

Di-Giovanni F, Kevan PG (1991) Factors affecting pollen dynamics and its importance to pollen contamination: a review. Can J For Res 21:1155–1170

Elant-Johnson, RC (1971) Probability models and Statistical methods in genetics. Wiley, New York

El-Kassaby YA, Ritland K (1986) Low levels of pollen contamination in a Douglas-fir seed orchard as detected by allozyme markers. Silvae Genetica 35:224–229

Epperson BK, Allard RW (1987) Linkage disequilibrium between allozymes in natural populations of Lodgepole pine. Genetics 115:341–352

Friedman ST, Adams WT (1981) Genetic efficiency in loblolly pine seed orchards. In: Proc 10th South Forest Tree Improv Conf. Blacksburg, Virginia, pp 213–234

Hamrick JL, Schnabel A (1985) Understanding the genetic structure of plant populations: some old problems and a new approach. In: Gregorius HR (ed) Population genetics in forestry. Springer, Berlin Heidelberg New York, pp 50–70

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

Neale DB (1984) Population genetic structure of the Douglas-fir shelterwood regeneration in southwest Oregon. PhD thesis, Oregon State University, Corvallis

Schoen DJ, Stewart SC (1986) Variation in male reproductive investment and male reproductive success in white spruce. Evolution 40:1109–1121

Schoen DJ, Stewart SC (1987) Variation in male fertilities and pairwise mating probabilities in *Picea glauca*. Genetics 116:141–152

Searle, SR (1971) Linear models. Wiley, New York

Smith DB, Adams WT (1983) Measuring pollen contamination in clonal seed orchards with the aid of genetic markers. In: Proc 17th South Forest Tree Improv Conf. Athens, Georgia, USA, pp 68–77

Squillace AE, Long EM (1981) Proportion of pollen form non-orchard sources. In: E. C. Franklin (ed) Pollen management handbook. USDA Agriculture Handbook 587, Washington, D. C., pp 15–19

Yazdani R, Lindgren D, Stewart SC (1989) Gene dispersion within a population of *Pinus sylvestris*. Scand J For Res 4:295–306