Methods in Ecology and Evolution

Methods in Ecology and Evolution 2011, **2**, 349–361 doi: 10.1111/j.2041-210X.2010.00085.x

Estimating the variance of male fecundity from genotypes of progeny arrays: evaluation of the Bayesian forward approach

Etienne K. Klein[1,2]*, **Florence H. Carpentier**[1] **and Sylvie Oddou-Muratorio**[2]

[1]*INRA, UR 546, Biostatistique et Processus Spatiaux, F-84000 Avignon, France; and* [2]*INRA, UR 629, Ecologie des Forêts Méditerranéennes, F-84000 Avignon, France*

Summary

1. Characterizing fine scale mating patterns in plant populations makes it possible to investigate genetic drift, gene flow and selection gradients at a contemporary time scale. Molecular markers are valuable tools for this type of analysis, and numerous statistical methods have been developed to make the best of the information they can provide. In particular, we recently proposed a Bayesian approach based on a paternity analysis wherein we estimate jointly the variance in male fecundity and the pollen dispersal kernel.

2. Here, we use simulated data sets to investigate the accuracy of the Bayesian approach compared to (i) classical maximum likelihood approaches (e.g. Neighbourhood model) that ignore variance in male fecundity or explain it through a few covariates and (ii) indirect methods (KinDist and Two-Gener) that integrate the variance in fecundity in an 'effective population density'.

3. The Bayesian estimates correctly considered the over-dispersion resulting from the variance in fecundity, resulting in wider but more accurate confidence intervals, in particular in high-density populations. The maximum likelihood methods resulted in confidence intervals with low coverage probabilities and in widespread false-positive tests when testing the effect of covariates on male fecundity.

4. Estimated individual fecundities and estimated empirical variance in fecundity were robust to the distribution assumed for the individual random fecundities (log-normal or Gamma). In contrast, the theoretical variance estimate critically depended on the assumed distribution.

5. The indirect methods provided much more variable estimators, as expected because they use less information about pollen sources and consider the molecular information only through genetic structure indices.

6. Disentangling the fecundity from the spatial effects in paternity analyses is necessary when studying selection *in natura* and or when addressing the effects of spatial distribution on effective gene flow. The Bayesian approach studied here successfully accounts for the variance in fecundity when a large fraction of it is not explained by the studied covariates. The Mixed Effect Mating Model computer program introduced here is devoted to its implementation.

Key-words: long-distance dispersal, male reproductive success, mating system, microsatellite markers, mixed effects mating models, paternity analysis, pollen dispersal, progeny array

Introduction

Genetic drift and gene flow between and within populations are two main evolutionary forces that interact with selection to determine the potential for adaptation. Evaluating the relative weights of these forces is especially important when investigating the fate of populations confronted with environmental changes (e.g. global warming or landscape fragmentation) and in the development of management strategies to temper the impacts of these environmental shifts (Ellstrand 1992; Savolainen, Pyhajarvi & Knurr 2007). Numerous recent studies have been published that characterize the mating systems, variances of reproductive success and gene flow at the instantaneous time scale, named contemporary approaches (Sork *et al.* 1999; Bacles *et al.* 2005).

*Correspondence author. E-mail: etienne.klein@avignon.inra.fr
Correspondence site: http://www.respond2articles.com/MEE/

To measure genetic drift and departure from random mating in plant populations, early studies used neutral genetic markers and paternity assignment to evaluate the male reproductive success (i.e. male fertility) of all pollen donors in an experimental plot (Devlin & Ellstrand 1990; Smouse & Meagher 1994). The inter-individual variance of fertility is then directly linked to the effective size of the population, N_{ep}. Simple exclusion (Chakraborty, Meagher & Smouse 1988), categorical paternity assignment (Meagher 1986; Marshall *et al.* 1998) and fractional paternity assignment (Devlin, Roeder & Ellstrand 1988; Nielsen *et al.* 2001) all resulted in large variance in male reproductive success only partially explained by phenotypic or micro-environmental variables (Smouse, Meagher & Kobak 1999; Smouse & Sork 2004). For instance, size, reproductive dominance and pollen production were often significantly related to fertility.

Among other variables, the distance between a pollen donor and a mother plant plays a specific role in determining their mating probability (Adams, Griffin & Moran 1992; Streiff *et al.* 1999). First, distance was almost always found to have a very significant effect on mating probability (Smouse & Sork 2004). Second, pollen dispersal limited by distance both determines gene flow at long distance and contributes to genetic drift and mating patterns at a very local scale (Garcia *et al.* 2005). Finally, the spatial pattern of pollen donors is most easily modified by human management and thus can be used as a lever to modify patterns of gene flow and genetic drift (Fernandez & Gonzalez-Martinez 2009).

Recent work has attempted to characterize the spatial component of mating patterns (Broquet & Petit 2009). The first step is to precisely estimate the pollen dispersal kernel, i.e. the probability density function describing the probability for a pollen grain emitted at a central point to pollinate an ovule at any position in space (Klein, Lavigne & Gouyon 2006). Numerous studies have characterized both the scale of pollen dispersal and the shape of the dispersal kernel, and thus the intensity of long-distance pollen dispersal (Austerlitz *et al.* 2004; Burczyk, Lewandowski & Chalupka 2004; Oddou-Muratorio, Klein & Austerlitz 2005; Robledo-Arnuncio & Gil 2005; Goto *et al.* 2006; Shimatani *et al.* 2007). Some studies have also compared several families of dispersal functions with different shapes (exponential-power vs. power-law tails) (Austerlitz *et al.* 2004; Klein, Lavigne & Gouyon 2006).

Two types of statistical approaches have been used to estimate pollen dispersal kernels in recent years: (i) the indirect approaches such as TwoGener and Kindist that rely on genetic distance or similarity indices among pollen pools sampled by pairs of mother trees (Austerlitz & Smouse 2002; Robledo-Arnuncio, Austerlitz & Smouse 2006) and (ii) spatially explicit mating models that use a maximum likelihood (ML) approach to integrate parentage, spatial and fecundity information (Burczyk *et al.* 2002; Oddou-Muratorio, Klein & Austerlitz 2005).

A second step in the analysis of the spatial component of mating patterns consists of getting rid of the effect of the relative positions of pollen donors and mother plants to estimate male fecundities (i.e. the amount of pollen released before dispersal) rather than male fertilities (or male reproductive success, i.e. the amount of offspring actually fertilized by a given pollen donor). This was partially achieved in mating models that estimated selection gradients (i.e. fixed effects of studied covariates on individual fecundity) in a spatially explicit context [e.g. the NEIGHBOURHOOD model, (Burczyk *et al.* 2002)]. Further, Klein, Desassis & Oddou-Muratorio (2008) attempted to estimate all of the individual fecundities of the pollen donors present in a study plot, and thus to estimate the entire variance in fecundity rather than the small part explained by the studied covariates. Because the variance in fecundity is related to the effective density of pollen donors d_{ep}, the results obtained with this approach can be compared with those from the indirect approaches that directly estimate this parameter (Robledo-Arnuncio, Austerlitz & Smouse 2007). Klein, Desassis & Oddou-Muratorio (2008) used a Bayesian approach relying on a Monte-Carlo Markov Chain to estimate the individual fecundities and the effective density of pollen donors d_{ep} and applied it to the *Sorbus torminalis* (ST) data set previously analysed with several classical approaches (Oddou-Muratorio *et al.* 2003; Austerlitz *et al.* 2004; Oddou-Muratorio, Klein & Austerlitz 2005, 2006). However, no investigation of the performance of the approach on simulated data sets was provided, although this would have helped to evaluate the potential benefits of the approach. In addition, they did not provide a computer program for researchers who might be eager to apply this approach to other data sets.

In this study, our goals were (i) to evaluate the accuracy and robustness of the estimates of the variance in fecundity obtained from the Bayesian approach when compared to other available methods, (ii) to investigate the effect of neglecting a source of variation of fecundity on the results of likelihood ratio tests (iii) to understand how variance in fertility differs from variance in fecundity and (iv) to introduce a computer program to apply easily the Bayesian statistical analyses presented here.

Materials and methods

SIMULATED DESIGNS – RANDOM DISTRIBUTIONS

Scenario 'low density (LD)'

We simulated 50 study populations distributed following a Poisson distribution with density 0·35 ha^{-1} in a disk with radius 13·49 (× 100 m) placed in the centre of a total population distributed in a disk with radius 134·9 following the same distribution. These radii were chosen to provide on average 200 trees inside the study area and 19 800 additional trees outside the study area (example given in Fig. 1a). Sixty mother trees were randomly sampled within the study populations.

Scenario 'high density (HD)'

We also simulated 50 study populations distributed following a Poisson distribution with density 3·5 ha^{-1} in a rectangle with dimensions 5 × 11·4 placed in the centre of a total population distributed in a disk with radius 50 following the same distribution. These dimensions provide on average 200 trees inside the study area and 27 290 additional trees outside the study area (example given in Fig. 1b). Sixty mother

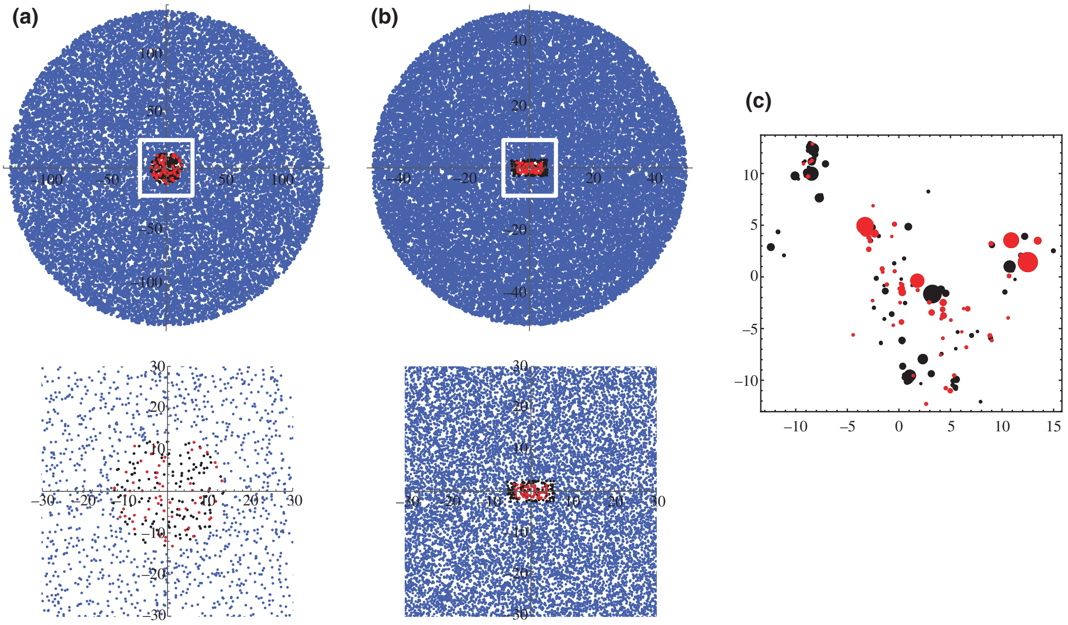

Fig. 1. Typical spatial configurations investigated in the three scenarios simulated: (a) Low density; (b) High density and (c) *Sorbus torminalis*. Bottom figures in (a) and (b) represent zooms on the white central squares. Blue dots are pollen donors not sampled, black dots are sampled pollen donors and red dots represent mother trees where seeds are collected.

trees were randomly sampled in the core of the study area, i.e. in the rectangle with dimensions 3·7 × 8·1 centred in the study rectangle.

SIMULATED DESIGNS – CLUSTERED DISTRIBUTIONS

Scenario 'Sorbus torminalis *(ST)'*

In 100 additional simulations, we used the actual positions of the 172 reproductive ST trees studied by Oddou-Muratorio, Klein & Austerlitz (2005). This population has a density of 0·35 ha^{-1} and is distributed in clusters of ~10 individuals on average in a ~100-m radius. In these simulations, we did not simulate trees outside the study plot but fixed the immigration rate at a constant value $m = 0{\cdot}4$.

SIMULATING MATING EVENTS

In simulations LD and HD, all trees were randomly assigned a genotype at six microsatellite loci with 6–24 alleles per locus and given allelic frequencies (see Data S1). These loci were MSS1, MMS5, MSS6, MSS9, MSS13, MSS16 with allelic frequencies as presented in Oddou-Muratorio *et al.* (2001). To draw the adult genotypes, we assumed linkage equilibrium among all loci, absence of inbreeding at all loci and no spatial genetic structure. The theoretical exclusion probability for this genetic system is 0·987. In simulations ST, the actual genotypes of the reproductive trees were used. They show spatial genetic structure with significant average kinship coefficients up to ~300 m (Oddou-Muratorio *et al.* 2004).

For each simulation *r* and tree *k*, we drew a fecundity value, $F_{r,k}$, in a log-normal distribution of mean 1 and variance $\sum_r^2 = e^{\sigma_r^2} - 1$. We then computed the composition of the pollen pools over the sampled mother trees as:

$$\pi_{r,jk} = \frac{F_{r,k} f(d_{jk};\delta,b)}{\sum\limits_{l:fathers} F_{r,l} f(d_{jl};\delta,b)} \quad \text{and } \pi_{r,jj} = 0 \qquad \text{eqn 1}$$

where π_{jk} is the proportion of pollen grains originating from the known father tree *k* in the pollen pool of mother tree *j*. *f* is the dispersal kernel with parameters δ (mean dispersal distance, scale parameter) and *b* (shape parameter), and d_{jk} is the distance between the mother tree *j* and the father tree *k*. We chose an exponential-power dispersal kernel (e.g. Klein, Desassis & Oddou-Muratorio 2008) with $\delta = 7{\cdot}5$ (× 100 m) and $b = 0{\cdot}3$:

$$f(x,y;\delta,b) = \frac{b\Gamma(3/b)^2}{2\pi\delta^2\Gamma(2/b)^3}\exp\left\lfloor\left(\frac{\Gamma(3/b)d}{\Gamma(2/b)\delta}\right)^2\right\rfloor \qquad \text{eqn 2}$$

with $d = \sqrt{(x_j - x_k)^2 + (y_j - y_k)^2}$ and Γ the classical gamma function.

For each of 1075 seeds sampled from the 60 mother trees (2–27 seeds per tree, following the same distribution as in the ST study), we then drew its father following the probabilities $\{\pi_{jk}\}_{k\ =\ 1\ldots}$ respecting independence among seeds. Knowing the mother and the father of each seed and their genotypes, we drew the genotype of the seed using Mendelian rules and independence among loci.

For scenarios LD and HD, we assume that no tree exists outside the simulated population ($m = 0$) and that selfing never occurs ($s = 0$). For the scenario ST where we did not simulate individually the external trees, we used the following algorithm for each seed *o* at each simulation *r*: with probability *m* (= 0·4) the pollen grain originated outside the site with a genotype drawn from the allelic frequencies. With probability *s* (= 0·02), the seed originated from a selfing event and the paternal gamete was drawn from the genotype of the mother tree using classical segregation probabilities, and with probability $(1 - m - s)$ (= 0·58) we drew the father from among the 171 known trees, apart from the mother, using the pollen pool composition $\{\pi_{jk}\}_{k\ =\ 1,\ldots,171}$ and the paternal gamete from the genotype of the retained father. A maternal gamete was drawn from the genotype of the mother tree and associated with the paternal gamete to provide the diploid genotype.

The first parameter of interest is the *variance in fecundity* Σ_r^2. This is the theoretical variance of the distribution of individual fecundities. In our simulations, this parameter is fixed. This parameter is related to the *theoretical ratio*

$$\left(\frac{d_{\text{obs}}}{d_{\text{ep}}}\right)_r^{th} = \Sigma_r^2 + 1 \qquad \text{eqn 3}$$

The second quantity of interest is the *empirical variance* S_r^2 of the actual fecundities of the trees within the study site. It is also unknown and varies from site to site. For each simulation r, it is computed from the individual fecundities $F_{r,k}$ as

$$S_r^2 = \frac{1}{n-1}\sum_k (F_{r,k} - \overline{F_r})^2, \qquad \text{eqn 4}$$

where n is the number of father trees inside the study site and $\overline{F_r}$ is the average of fecundities $F_{r,k}$. This parameter is related to the *empirical ratio*

$$\left(\frac{d_{\text{obs}}}{d_{\text{ep}}}\right)_r^{emp} = S_r^2 + 1 \qquad \text{eqn 5}$$

We also define the *variance in fertility* over the sampled seeds as

$$SF_r^2 = \frac{1}{n-1}\sum_k (N_{r,k} - \overline{N}_r)^2, \qquad \text{eqn 6}$$

where $N_{r,k}$ is the number of sampled seeds actually fathered by the tree k and $\overline{F}_r$ is the average of the fertilities $N_{r,k}$.

In scenarios LD and HD, we simulated one data set for each of Σ_r^2 varying from 0 to 10 by steps of 0·2, resulting in 50 data sets. In the scenario ST, we simulated one data set for each of Σ_r^2 varying from 0 to 10 by steps of 0·1, resulting in 100 data sets.

SIMULATING COVARIATES NOT AFFECTING FECUNDITY

We investigated how variance of fecundity affects the probability that a likelihood ratio test incorrectly detects a significant effect of a covariate on fecundity (Type I error rate). To this goal, in the LD and HD scenarios, for each adult tree inside the study area we independently drew two discrete covariates COV1 and COV2 with probabilities (0·9, 0·16, 0·25, 0·25, 0·16, 0·9) associated with six modalities (A, B, C, D, E, F). In the ST scenario, there were three covariates associated with each reproductive tree: flowering intensity (COV1), local density (COV2) and diameter class (COV3) with 4, 5 and 6 modalities respectively. None of these variables had any effect on fecundity or fertility in the simulated data sets.

ESTIMATION OF THE DISPERSAL PARAMETERS AND VARIANCE IN FECUNDITY

For each data set, we used each of the following approaches to estimate dispersal parameters and variance in fecundity: the ML approach based on a mating model (Burczyk *et al.* 2002; Oddou-Muratorio, Klein & Austerlitz 2005) assuming the same fecundity for all trees, the same maximum likelihood approach using the covariates (MLCov) as proxies for fecundity, (BayLN) the Bayesian approach modelling fecundity through an individual random effect log-normally distributed (Klein, Desassis & Oddou-Muratorio 2008), (BayG) the Bayesian approach with a gamma distribution of individual fecundities and (KD+TG) the Kindist and TwoGener approaches associated as suggested in Poldisp (Robledo-Arnuncio, Austerlitz & Smouse 2007). For the MLCov approach, we used the two covariates COV1 and COV2 in the LD and HD scenarios and the three actual covariates COV1-3 in the ST scenario.

Using a Gamma distribution (BayG) in the estimation procedure whereas the data were simulated using a log-normal distribution for the fecundities aims at investigating the robustness of the estimates regarding the distribution chosen. Fitting the effects of covariates that were not used in the simulations (MLCov) provides an evaluation of the Type I error rate of the likelihood ratio tests.

Maximum likelihood and mating model – In the approaches ML and MLCov, we used the likelihood function previously defined (Adams & Birkes 1991; Burczyk *et al.* 2002; Oddou-Muratorio, Klein & Austerlitz 2005), associated with the set of genotypes $g = (g_o)_{\text{o:offsprings}}$ of the sampled seeds:

$$L(\mathbf{g}|\alpha_{\text{COV}}, \delta, b, s, m) = \prod_{\text{o:offspring}} \Bigg[sT(g_o|g_{j_o}, g_{j_o}) + mT(g_o|g_{j_o}, BAF) + (1-s-m)\sum_{\text{k:father}} \pi_{j_o k} T(g_o|g_{j_o}, g_k) \Bigg] \qquad \text{eqn 7}$$

where π_{jk} is the composition of the pollen pool (eqn 1) which depends on the dispersal parameters only when using ML approach (see eqn 5 in Oddou-Muratorio, Klein & Austerlitz 2005). When using the MLCov approach, π_{jk} depends both on the dispersal parameters and on the parameters of the covariates (α_{COV} in the approach MLCov, see eqn 4 in Oddou-Muratorio, Klein & Austerlitz 2005). $T(g_o|g_{j_o}, X)$ is the Mendelian segregation probability (Meagher 1986) of the offspring genotype (g_o) given the genotype of the mother (g_{j_o}) and X, where X corresponds (i) to the genotype of the mother in the case of self-fertilization (ii) to the allelic frequencies in the pollen pool external to the neighbourhood (*BAF*) in the case of outcrossing with a non-sampled father tree or (iii) to the genotype of the considered father tree (g_k) in the case of outcrossing with a sampled male k.

Dispersal parameters (δ, b), mating parameters (s, m) and effects of covariates (α_{COV}, only in the MLCov case) were estimated by maximizing the likelihood function using Mathematica 7·1. In the MLCov approach, we computed the variance in fecundity and the ratio $d_{\text{obs}}/d_{\text{ep}}$ from the estimated α_{COV} following eqn 8 in (Oddou-Muratorio, Klein & Austerlitz 2005).

We tested the significance of all of the effects of the covariates globally using a likelihood ratio test that compared the likelihood reached in the MLCov with that reached in the ML approach (Johnson & Omland 2004). We did not investigate whether the effects of the covariates (α_{COV}) were correctly estimated but instead focused on the probability to conclude that the covariates significantly determine fecundity although they actually did not.

Bayesian estimation and random individual fecundity – We used the approach developed in Klein, Desassis & Oddou-Muratorio (2008) to estimate (i) the dispersal parameters (δ, b) and the mating system parameters (s, m), (ii) all individual relative fecundities F_k and (ii) the ratio $d_{\text{obs}}/d_{\text{ep}}$, which measures the variance in fecundity. The approach relies on the likelihood for the set of genotypes $g = (g_o)_{\text{o:offspring}}$ of the sampled seeds

$$L(\mathbf{g}|\mathbf{F}, \sigma^2, \delta, b, s, m) = \prod_{\textit{offspring}} \Bigg[sT(g_o|g_{j_o}, g_{j_o}) + mT(g_o|g_{j_o}, BAF) + (1-s-m)\sum_{\textit{k:father}} \pi_{j_o k} T(g_o|g_{j_o}, g_k) \Bigg], \qquad \text{eqn 8}$$

where the composition of the pollen pools π_{jk} now depends on the dispersal parameters and on the individual fecundities $F = \{F_k\}_{k=1\ldots}$ as given by eqn 1. The transition probabilities $T(g_o|g_{j_o}, X)$ are defined as above.

We used the same prior distributions and the same proposal distributions as in Klein, Desassis & Oddou-Muratorio (2008). We computed the posterior distributions for the parameters $\sum_r^2, \delta, b, m, s$ using a Mote Carlo Markov Chain (MCMC) of 50 000 steps after a 5000 step burn-in period. For all parameters, we computed the posterior mean and median and the 95%-credibility interval from the 50 000 steps of the MCMC. We initiated the Markov chain with $\left(\sum_{r,10}^2, \delta_0, b_0, m_0, s_0\right) = \left(e^2 - 1, 50, 1, 0 \cdot 5, 0 \cdot 1\right)$ and $\mathrm{F}_0 = (1,\ldots,1)$.

For each simulation r, the MCMC provided a posterior distribution for the parameter$\sum_r^2$ and the *estimated theoretical variance* was obtained as the posterior mean and posterior median

$$\hat{\Sigma}_r^2 = \frac{1}{T} \sum_{t=1,\ldots,T} \Sigma_r^{2(t)} \text{ and } \tilde{\Sigma}_r^2 = \text{median}\left(\left\{\Sigma_r^{2(t)}\right\}_{t=1,\ldots,T}\right), \qquad \text{eqn 9}$$

where T is the number of iterations retained. The associated ratio $d_{\text{obs}}/d_{\text{ep}}$ can be computed at each iteration t,

$$\left(\frac{d_{\text{obs}}}{d_{\text{ep}}}\right)_r^{(t)} = \sum\nolimits_r^{2(+)} + 1 \qquad \text{eqn 10}$$

and the *estimated theoretical ratio* $d_{\text{obs}}/d_{\text{ep}}$ was obtained as the posterior mean and the posterior median

$$\left(\frac{d_{\text{obs}}}{d_{\text{ep}}}\right)_r^{\wedge} \text{ and } \left(\frac{d_{\text{obs}}}{d_{\text{ep}}}\right)_r^{\sim} \qquad \text{eqn 11}$$

The posterior distribution and 95% credibility intervals were obtained by computing the 2·5% and 97·5-quantiles from the 50 000 retained values of the parameter in the MCMC.

Using the individual fecundities every 20 iterations,$F_r^{(t)}$,we also computed the posterior distribution for the fecundity of each individual k, $F_{r,k}$. The *estimated individual fecundities* were then obtained as the posterior mean:

$$\hat{F}_{r,k} = \frac{1}{T} \sum_t F_{r,k}^{(t)}, \qquad \text{eqn 12}$$

where T is the number of iterations retained.

We related these estimated fecundities to the actual fecundities by computing the coefficient of determination, R^2, for the log-log regression of $\{\log(\hat{F}_{r,k})\}_{k=1,\ldots,n}$ over $\{\log(F_{r,k})\}_{k=1,\ldots,n}$.

We also computed the *estimated empirical variance* of fecundities as

$$\hat{S}_r^2 = \frac{1}{n-1} \sum_k \left(\hat{F}_{r,k} - \overline{\hat{F}}\right)^2, \qquad \textit{eqn } 13$$

and the *estimated empirical ratio* $d_{\text{obs}}/d_{\text{ep}}$ as

$$\left(\frac{d_{\text{obs}}}{d_{\text{ep}}}\right)_r^{\wedge,emp} = \frac{\hat{S}_r^2}{\hat{F}_r^2} + 1 \qquad \text{eqn 14}$$

Kindist + TwoGener estimation – We applied Kindist to all data sets to estimate the parameters a, b and δ of an exponential-power dispersal kernel. To recalibrate the pairwise kinship coefficients 'automatically' for each data set, we set the threshold distance by choosing the maximum distance above which 50 pairwise distances are found. This led to threshold distances of 23·0 (× 100 m) and 6·8 on average for the LD and HD simulations and to the distance 23·9 for all the ST simulations.

The effective density d_{ep} was then estimated by applying the pairwise TwoGener with an exponential power dispersal kernel to all the data sets. The dispersal parameters a and b were fixed to the values previously estimated by Kindist and only the density was estimated. The true density (0·35 or 3·5) was used as initial value.

ESTIMATION SOFTWARE

The Bayesian analyses were achieved using the software Mixed Effect Mating Model (MEMM). This software implements the method presented here and in Klein, Desassis & Oddou-Muratorio (2008) to estimate the dispersal and mating parameters and the theoretical and empirical variance in male fecundity. Log-normal and gamma distributions for the random individual fecundities can be used.

Mixed Effect Mating Model is available at http://memm.biosp.org together with a manual and examples of input data files. Versions for MS Windows, MacOS X (intel) and linux are available. The C++ code is also available from E Klein upon request for users wishing to compile it on their own computer. The input and output files are text files, and functions to graphically plot the results are provided which run on R (CRAN project).

Results

ESTIMATES OF THE DISPERSAL KERNEL AND MATING SYSTEM PARAMETERS

The Bayesian approach BayLN provided estimates of the dispersal parameters independent of the variance in fecundity $d_{\text{obs}}/d_{\text{ep}}$ for the three scenarios (e.g. Fig. 2). The dispersal parameter estimates were slightly biased towards more long-distance dispersal: (i) the mean dispersal distance δ was biased upward and (ii) the shape parameter was biased downward, i.e. towards fatter-tailed dispersal kernels (Table 1). For the mean dispersal distance, the estimates based on the posterior median were less biased than those resulting from the posterior mean (geometric mean of the estimates = 8·18 vs. 8·85 for ST; 8·68 vs. 11·15 for LD; 12·3 vs. 24·2 for HD). They also had a smaller standard deviation (not shown). Using a gamma distribution (BayG) for the fecundities instead of the log-normal distribution led to a slight increase in the bias and standard deviation (geometric mean of the posterior median estimates = 8·54 vs. 8·18 for ST; 8·93 vs. 8·68 for LD and 21·2 vs. 12·3 for HD).

The 95%-credibility intervals computed in the Bayesian approach from the posterior distribution were globally accurate, with error rates between 0% (parameter b for BayLN in scenario ST) and 16% (parameter b for BayG in scenario LD) where we expected 5%. Finally, scenario HD provided wider credibility intervals, especially for the mean dispersal distance δ but also for the shape parameter b. This scenario was also associated with higher bias and greater variance of the estimates (Table 1).

The maximum likelihood estimators (methods ML and MLcov in Table 1) were almost unbiased for the dispersal parameters δ (means for δ estimates were 7·95, 7·89 and 7·70 instead of 7·5 for ML in scenarios ST, LD and HD) and b (means for b estimates were 0·35, 0·34 and 0·33 instead of 0·3 for ML in scenarios ST, LD and HD). The 95%-likelihood-profile confidence intervals for these parameters were narrower

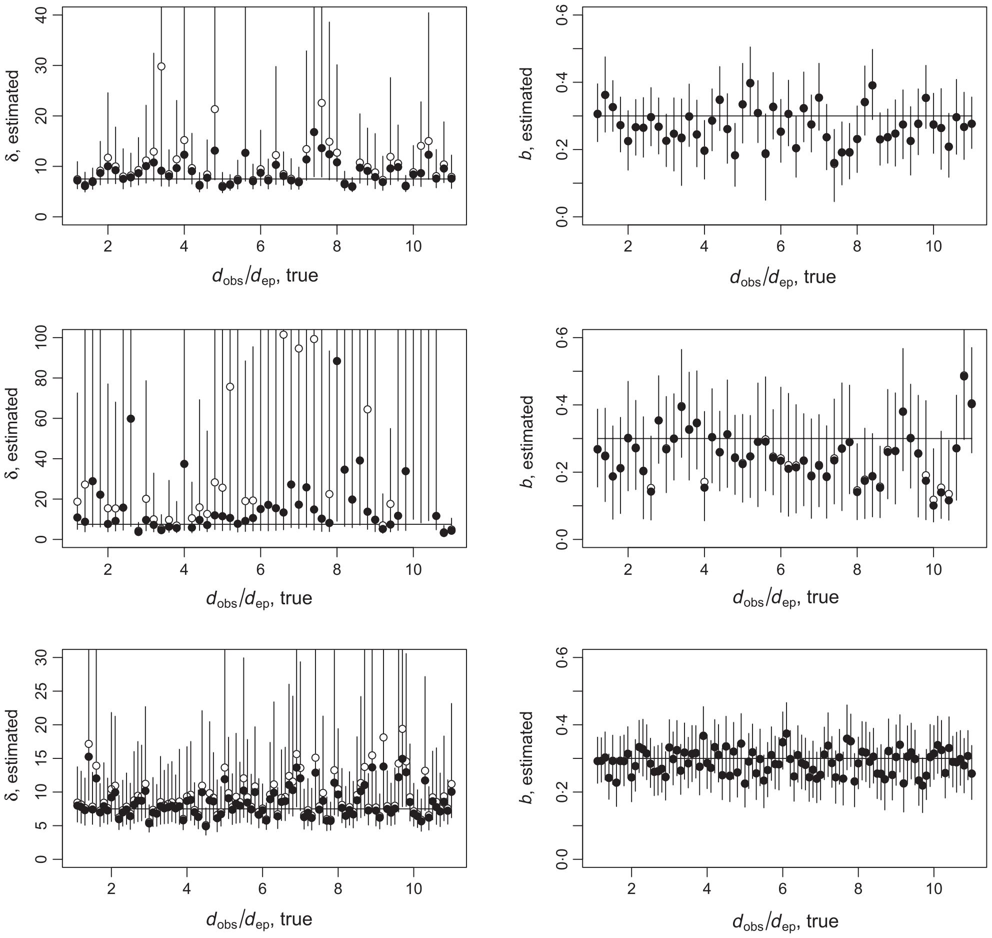

Fig. 2. Values estimated for the dispersal parameters δ (scale parameter, left column) and *b* (shape parameter, right column) using the BayLN approach. Scenarios low density, high density and *Sorbus torminalis* are presented from top to bottom. *x*-axis represents the fixed theoretical ratio d_{obs}/d_{ep} (eqn 3). Full dots are the estimates based on the posterior median, empty dots are for the posterior mean, and bars represent the 95% credibility interval. The black line represents the true value of the parameter.

than the 95% credibility interval provided by BayLN. In fact they were too narrow, as the true value for the parameters $\delta = 7{\cdot}5$ and $b = 0{\cdot}3$ was not included in the confidence intervals in more than 5% of the simulations (16% for δ in the LD case to 42% for *b* in the ST case). Testing ($\delta = 7{\cdot}5$, $b = 0{\cdot}3$) simultaneously using a 2 DF likelihood ratio test resulted in the rejection of the true values even more often (72%, 58% and 26% of the simulations in the ST, LD and HD scenarios). The estimates obtained with (MLCov) or without (ML) considering the covariates for fecundity had similar properties (Table 1): considering inadequate covariates acting on fecundity did not affect the estimates of the dispersal parameters.

Furthermore, contrary to the Bayesian estimates, ML estimates of the dispersal parameters were affected by an increase in variance in fecundity. First, the bias in the estimation of δ increased with d_{obs}/d_{ep} (e.g. for the ST case: geometric mean = 7·39 for δ when d_{obs}/d_{ep} is in (1,6); geometric mean = 8·55 for δ when d_{obs}/d_{ep} in (6,11); geometric means significantly different, $P = 0{\cdot}01$). Second, the accuracy of the confidence intervals decreased when d_{obs}/d_{ep} increased: in the ST scenario, if 72 simulations over 100 found the true value (7·5, 0·3) outside of the 95%-confidence area, this type I error was 25% over the 20 simulations with $1 < d_{obs}/d_{ep} \leq 3$ and 85% over the 80 simulations with $3 < d_{obs}/d_{ep} \leq 11$.

Finally, the Kindist estimates showed a moderate bias for the mean dispersal distance, but the parameter *b* was overestimated (means of the *b* estimates were 0·39, 0·48 and 0·94 instead of 0·3 in scenarios ST, LD and HD). However, several simulations led to extreme values for δ and/or *b*. Surprisingly, the method seems more affected by HD than by a clustered distribution of pollen donors (more bias in the HD scenario than in the ST scenario, Table 1). The Kindist estimates were mostly characterized by a large variance apparently independent of the variance in fecundity (Figure S1).

ESTIMATES OF THE VARIANCE OF MALE FECUNDITY

The estimated theoretical variance (eqn 9) provided by the Bayesian approach BayLN accurately estimated the variance in fecundity (Fig. 3, Table 2). The estimated theoretical ratio d_{obs}/d_{ep} (eqn 11) had low bias (Fig. 1; mean relative bias of

Table 1. Mean values for the dispersal parameters estimated from the five methods in the three scenarios *Sorbus torminalis* (ST), low density (LD) and high density (HD) scenarios and obtained over all values of the variance in fecundity. For each parameter and each method, we provide the geometric mean of the estimates, the mean confidence/credibility interval and the percentage of cases where the true value of the parameter was outside the 95%-confidence interval. For Bayesian methods BayLN and BayG, we provide in each cell the performance of the posterior mean (left) and the posterior median (right)

Scenario	Method	Mean dispersal distance, $\delta = 7{\cdot}5$	Shape parameter, $b = 0{\cdot}3$	Migration rate, m*	Selfing rate, s
ST	BayLN	8·85/8·18 (5·5; 16·0) 6%	0·29/0·29 (0·21; 0·37) 0%	0·45/0·45 (0·42; 0·48) 11%	0·011/0·011 (0·005; 0·019) 62%
	BayG	9·35/8·54 (5·62; 17·31) 7%	0·29/0·28 (0·20; 0·37) 3%	0·45/0·45 (0·42; 0·48) 8%	0·011/0·011 (0·005; 0·019) 59%
	ML	7·95 (5·8; 12·4) 23%	0·35 (0·28; 0·42) 42%	0·45 (0·42; 0·48) 13%	0·011 (0·006; 0·019) 50%
	MLCov	8·05 (5·8; 12·8) 25%	0·34 (0·27; 0·41) 38%	0·45 (0·42; 0·48) 12%	0·011 (0·005; 0·019) 50%
	KinDist	14·3	0·39	–	–
LD	BayLN	11·15/8·68 (6·12; 20·6) 4%	0·266/0·265 (0·16; 0·36) 14%	0·31/0·31 (0·29; 0·34)	$<10^{-8}$ (0; $1{\cdot}10^{-7}$)
	BayG	11·45/8·93 (6·2; 23·4) 10%	0·264/0·264 (0·16; 0·36) 16%	0·31/0·31 (0·29; 0·34)	$<10^{-8}$ (0; $1{\cdot}10^{-7}$)
	ML	7·89 (6·03; 11·87) 16%	0·34 (0·26; 0·43) 36%	0·31 (0·29; 0·34) 64%%	–
	MLCov	7·88 (6·09; 11·89) 10%	0·34 (0·25; 0·42) 30%	0·31 (0·29; 0·34) 64%	–
	KinDist	9·24	0·48	–	–
HD	BayLN	24·2/12·3 (4·5; 116) 8%	0·25/0·25 (0·13; 0·41) 6%	0·54/0·54 (0·51; 0·57)	$<10^{-6}$ (0; $7{\cdot}10^{-5)}$
	BayG	70·0/21·2 (4·9; 383) 14%	0·232/0·221 (0·11; 0·39) 12%	0·54/0·54 (0·51; 0·57)	$<10^{-6}$ (0; $2{\cdot}10^{-6}$)
	ML†	7·70 (5·11; 28·6) 20%	0·33 (0·19; 0·48) 18%	0·54 (0·51; 0·57) 46%	–
	MLCov†	7·85 (4·21; 34·3) 14%	0·32 (0·18; 0·47) 10%	0·54 (0·51; 0·57) 44%	–
	KinDist‡	2·25	0·94	–	–

*The true value for m was 0·45 in the ST scenario. In the LD and HD scenarios, we computed the true value by averaging over the simulations the proportion of real paternities actually out of the study site. We found $m = 0{\cdot}32$ and 0·54 for LD and HD respectively.

†The simulation $d_{\mathrm{obs}}/d_{\mathrm{e}} = 10{\cdot}2$ was removed because the absence of convergence led to unrealistic values ($\delta > 10^5$ and $b < 0{\cdot}1$)

‡Three simulations were removed ($d_{\mathrm{obs}}/d_{\mathrm{e}} = 1{\cdot}2$, 1·8 and 10·8) because they led to unrealistic values ($d_{\mathrm{e}} > 1000$ and $\delta < 0{\cdot}1$)

–, not estimated.

14%, 8% and 2% for the posterior median in scenarios ST, LD and HD). Similar to the dispersal parameters, the posterior mean was more biased than the posterior median. The variance of the estimated theoretical ratio (eqn 11) and the skewness of the posterior distribution increased notably when the true variance in fecundity increases (Fig. 3). This explains the discrepancy between posterior means and posterior medians. This also results in wide confidence intervals for large variances in fecundity.

When we used a Gamma distribution in the estimation procedure (BayG), (although it was actually log-normally distributed, cf. M&M section) the theoretical variance was strongly under-estimated, in particular for ratios $d_{\mathrm{obs}}/d_{\mathrm{ep}} > 3$ (Fig. 3; mean relative biases ~ −45%). In fact, the BayG method rarely estimated values above to 4 for the ratio $d_{\mathrm{obs}}/d_{\mathrm{ep}}$ (Fig. 3).

Regarding the empirical variance of fecundity (S_r^2givenby eqn 4), the estimator $\hat{S}_r^2$ (eqn 13) based on estimated individual fecundities ($\hat{F}_{r,k}$ given by eqn 12) was more robust to the choice of the distribution (Gamma or LN) than the estimator $\hat{\Sigma}_r^2$. First, the estimated empirical variances obtained with the BayLN and BayG methods were quite close to each other, even if those obtained from the BayG were generally lower than those from BayLN (Fig. 4). This was confirmed by the strong correlation between the individual fecundities estimated in BayG and in BayLN (the average correlation coefficient was 0·985; 0·992 and 0·991 in scenarios ST, LD and HD). Second, the estimated empirical variances were closer to the true empirical variances than the estimated theoretical variances, even with the BayLN method (Fig. 4). This means that the individual fecundities were well estimated even when the distribution chosen to model them was wrong (gamma instead of log-normal). The average correlation between the true individual fecundities and the estimated individual fecundities supported this result (average correlation = 0·888, 0·921, 0·897 for the BayLN estimates and average correlation = 0·878, 0·916, 0·892 for the BayG estimates).

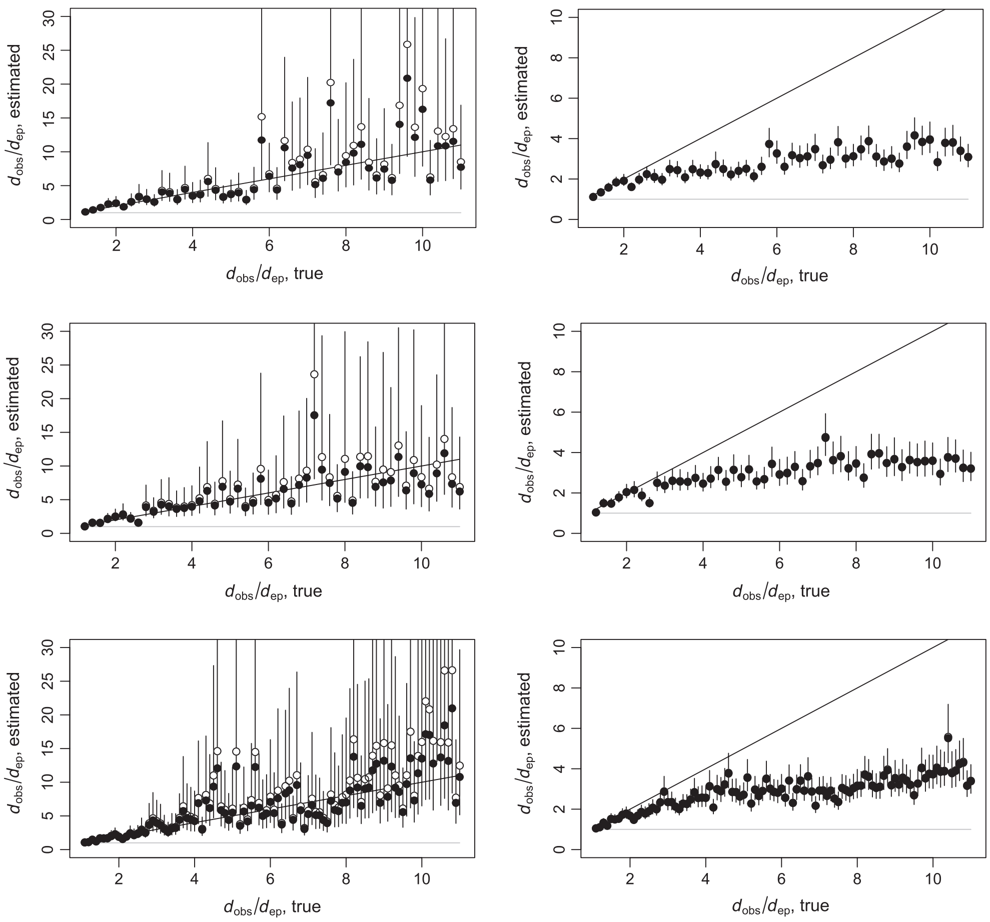

Fig. 3. Values estimated for the parameter measuring the variance in fecundity using the Bayesian approach. *x*-axis represents the true theoretical ratio d_{obs}/d_{ep} (eqn 3) and *y*-axis the estimated theoretical ratio (eqn 10). Scenarios low density, high density and *Sorbus torminalis* are presented from top to bottom. Left figures present parameters estimated using a log-normal random fecundity (BayLN) and right figures parameters estimated with a Gamma random fecundity (BayG). Full dots are the estimates based on the posterior median, empty dots are for the posterior mean, and bars represent the 95% credibility interval. The black line represents the diagonal 'estimated value = true value'.

Table 2. Variance in fecundity parameters estimated from the five methods. Mean relative bias and mean relative confidence interval are provided for d_{obs}/d_e. Geometric mean is provided for d_e estimates. Type I error rates for MLCov provide the percentage of simulated data sets for which the likelihood ratio 5%-test concluded in a significant effect of the covariates that actually had no effect on fecundity (we would thus expect values close to 5%)

Scenario	Method	d_{obs}/d_e	d_e (trees ha^{-1})*	Type I error rate
Sorbus torminalis	BayLN	+30%/+14% (−0·34; 1·88) 8%		
	BayG	−44%/−45% (−0·53; −0·33) 89%		
	MLCov	−71% max = 3·72		97%
	KinDist+ TwoGener	+290%	0·067 R^2 = 0·001	
Low density	BayLN	+19%/+8% (−0·32; 1·36) 8%		
	BayG	−46%/−46% (−0·54; −0·37) 88%		
	MLCov	−74% max = 1·63		96%
	KinDist + TwoGener	+31%	0·12 R^2 = 0·004	
High density	BayLN	+12%/+2% (−0·37; 1·24) 6%		
	BayG	−42%/−43% (−0·52; −0·30) 88%		
	MLCov[†]	−72% max = 2·04		94%
	KinDist + TwoGener[‡]	+13%	3·89 R^2 = 0·011	

*R^2 indicates the coefficient of determination for the regression of the estimated density d_e against the true d_{obs}/d_e.

[†] The simulation d_{obs}/d_e = 10·2 was removed because the absence of convergence led to unrealistic values ($\delta > 10^5$ and $b < 0{\cdot}1$)

[‡] Three simulations were removed (d_{obs}/d_e = 1·2, 1·8 and 10·8) because they led to unrealistic values ($d_e > 1000$ and $\delta < 0{\cdot}1$)

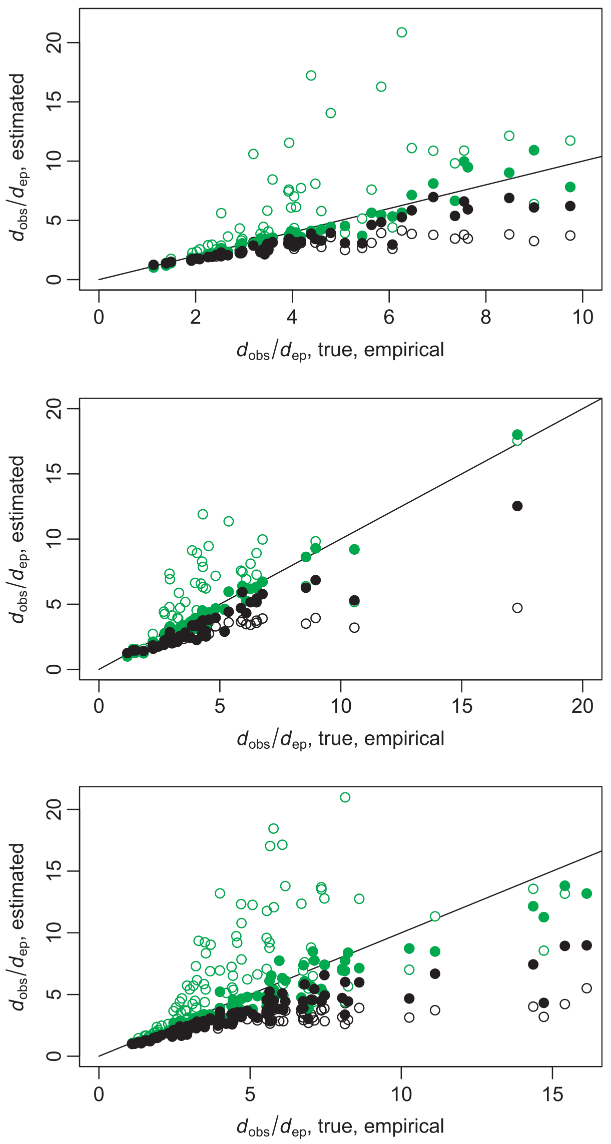

Fig. 4. Estimated values for the parameters measuring the variance in fecundity. True empirical ratio d_{obs}/d_{ep} (eqn 5) computed from the actual fecundity of the simulated putative fathers inside the study plot are plotted on the *x*-axis. On the *y*-axis, we plotted the estimated theoretical ratio d_{obs}/d_{ep} (eqn 10, empty marks) and the estimated empirical ratio d_{obs}/d_{ep} (eqn 13, full marks). Estimations obtained from a log-normal assumption for the distribution of the random individual fecundity are plotted in green. Estimations obtained from a gamma assumption are plotted in black. Scenarios low density, high density and *Sorbus torminalis* are presented from top to bottom.

As expected, the ML approach with covariates that were not actually related to fecundity (MLCov) was unable to estimate correctly the variance in fecundity (Table 2, mean relative biases ~ −70%). However, note that some unexpectedly high variances of fecundity were estimated for some particular simulations (estimated d_{obs}/d_{ep} of 3·72, 1·63 and 2·04 were obtained for scenario ST d_{obs}/d_{ep} = 9·7, scenario LD d_{obs}/d_{ep} = 9·6 and scenario HD d_{obs}/d_{ep} = 9·6).

Kindist and Two-Gener estimated values for d_{ep} that were in the correct order of magnitude (d_{ep} estimated at 0·067 and 0·12 on average for ST and LD where d_{obs} = 0·35 and thus d_{ep} = 0·35 to 0·032; d_{ep} estimated at 3·89 on average for HD where d_{obs} = 3·5 and thus d_{ep} = 3·5 to 0·32). However, among different simulations of the same scenario, the estimated values d_{ep} were not correlated with the true value d_{obs}/d_{ep} (R^2 = 0·001, 0·004 and 0·011, P = 0·08, 0·17 and 0·51 for ST, LD and HD).

RELATION BETWEEN MALE FECUNDITY AND MALE FERTILITY (I.E. MALE REPRODUCTIVE SUCCESS)

In our simulations, the average correlation between individual fecundities and expected individual fertilities was (only) 0·8, 0·89 and 0·87 in the three scenarios investigated. This correlation was the lowest in the ST scenario, where both LD and clustered distribution of trees were expected to provide greater stochasticity.

Assuming that all paternities could have been retrieved perfectly from a paternity analysis resulting in observed mating success equal to the number of progeny found for each pollen donor, we could be tempted to use these observed individual fertilities as estimates of individual fecundities. However, they provided worse estimates than the Bayesian approach: the average correlation between the individual fecundities and the observed individual fertilities was 0·75, 0·84 and 0·81 (in ST, LD and HD scenarios), whereas the average correlation with the BayLN estimate of fecundity was 0·89, 0·92 and 0·90 and the correlation with the BayG estimates was 0·88, 0·92 and 0·89.

NEGLECTING THE VARIANCE OF MALE FECUNDITY IN ML APPROACHES

As mentioned earlier, the confidence intervals for dispersal parameters were too narrow when d_{obs}/d_{ep} increased, leading to excessive rates of CI not containing the true value. Further-

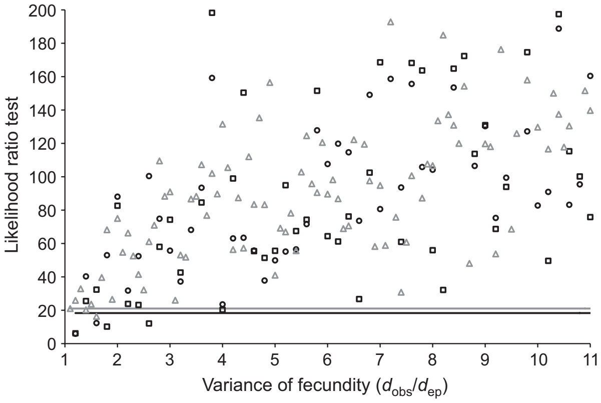

Fig. 5. Type I errors of the likelihood ratio test aiming at detecting significant effects of covariates on the fecundity. The values of the LRT for all simulated data sets are represented as a function of the theoretical ratio d_{obs}/d_{ep} (*x*-axis, eqn 3). All three scenarios are plotted [Squares: low density (LD); Circles: high density (HD); Triangles: *Sorbus torminalis* (ST)]. The horizontal lines represent the threshold values above which the covariates are considered as significant at the 5% level (Black: LD and HD; Grey: ST). 16 values of the LRT > 200 are not plotted on this scale.

more, the MLCov approach wrongly concluded that there were significant effects of the covariates in 192 simulations of 200 in all three scenarios (Fig. 5). The only eight simulations where the likelihood ratio test correctly concluded that there were non-significant effects of the covariates were all concentrated in the range $d_{\mathrm{obs}}/d_{\mathrm{ep}} < 3$.

Discussion

We ran two hundred simulations to evaluate the Bayesian estimation of the pollen dispersal kernel and the variance of male fecundity in a realistic context concerning the number of putative parents, the number of mother trees and sampled seeds and the exclusion power of the genetic system used. Although the statistical theory provides theoretical properties for large data sets (asymptotic results), it is also necessary to investigate the actual properties of our estimates for typical data sets. This is particularly true for Bayesian algorithms where several tuning parameters (prior distributions, proposal distribution, number of iterations of the MCMC and burn-in iterations) may affect the results of the estimation procedure.

Overall the Bayesian estimates performed well, providing quite accurate estimates of the dispersal parameters and of the variance in fecundity (low bias). However, the variance and the width of the confidence intervals of the Bayesian estimates of the variance in fecundity increases quickly with the true variance in fecundity. Estimating precisely this parameter when large differences of pollen production and pollen efficiency among trees exist could thus be difficult. Including both the major covariates determining male fecundity and a random individual effect is a way to consider random individual effects with smaller variance. The dispersal parameters estimates were not affected by the variance in fecundity over the wide range investigated here ($d_{\mathrm{obs}}/d_{\mathrm{ep}}$ from 1 to 11). This is a valuable property of the analysis method because variance in fecundity can be high under real experimental conditions. The classical ML estimates, by contrast, were notably affected by the variance in fecundity, with a bias increasing with $d_{\mathrm{obs}}/d_{\mathrm{ep}}$ and confidence intervals and likelihood ratio tests becoming less accurate as $d_{\mathrm{obs}}/d_{\mathrm{ep}}$ increased.

We also found that the estimate of the theoretical variance in fecundity (i.e. that of an infinite populations of trees where the whole distribution of fecundities would be represented) was sensitive to the assumed distribution of fecundity (here gamma or LN). This is comparable to estimated mean dispersal distances, where different dispersal kernels fitted on the same data set result in very different mean dispersal distances. Mean dispersal distance and variance of fecundity both depend strongly on the tails of the curves (i.e. dispersal kernels and distribution of fecundity) and thus on the extrapolation of the processes observed within the site to a broader range. This problem can be minimized by (i) using several curves and selecting the one that best fit the data (e.g. this can be carried out using Bayes factors to select the Gamma or LN in the Bayesian scheme here, e.g. Klein, Desassis & Oddou-Muratorio 2008) and (ii) building experiments covering the largest range as possible (i.e. larger spatial scale and better sampling of the distribution of male fecundities will provide better extrapolations). Finally, the drawback of extrapolation is avoided when focusing only on the empirical variance in fecundity (i.e. that among the actual trees present in the study site) instead of the variance of the population.

We showed here that the estimated empirical variance was much less sensitive to the distribution assumed for fecundity. However, this parameter leads too less general conclusions as the empirical variance in fecundity can vary notably from site to site and be notably different from the variance of the population (e.g. Fig. 4).

This study was more focused on the estimation of the variance in fecundity, and we did not fully investigate the questions about the estimation of the dispersal function. In particular, we did not analyse the data using a kernel family different from the exponential power family used to generate them. In traditional ML approaches, this results in inaccurate estimates of the dispersal parameters and the same result is expected from the Bayesian approach with random fecundity. Using several kernel families and selecting the best fitting function remains a critical step in the analysis of experimental data, especially because of a lack of fit owing to a misspecification of the dispersal kernel could subsequently be compensated by incorrect estimated individual fecundities and biased variance in fecundity. More intensive sampling (more mother trees widespread in the site) should limit this drawback. Further simulations should investigate quantitatively this aspect, but a general conclusion expected is that the Bayesian approach is better suited for large data sets.

This article is also an opportunity to present the MEMM computer program available for those wishing to estimate variance in fecundity and the dispersal function using the Bayesian method first proposed in (Klein, Desassis & Oddou-Muratorio 2008) and investigated here. The program presently available at http://memm.biosp.org provides Bayesian estimates for the dispersal parameters and the variance in fecundity assuming a log-normal or a gamma distribution for the random individual fecundities and an exponential power function for the dispersal kernel. Future versions will include more diverse dispersal functions as several studies have shown that the shape of the dispersal function strongly affects the estimated mating patterns (Klein, Lavigne & Gouyon 2006; Robledo-Arnuncio & Austerlitz 2006). It thus deserves careful analysis through the application of a wide range of possible dispersal tails (Austerlitz *et al.* 2004; Goto *et al.* 2006).

Another biological phenomenon crucial for determining the mating events among individuals is asynchronous flowering (Kang *et al.* 2003; Gérard *et al.* 2006). Including a temporal distance between individuals that governs the probability of mating is possible in mating models (Smouse & Sork 2004) but requires the measurement of flowering phenology for all of the pollen donors in the study site. These data are costly to gather even if they sometimes prove to more significantly affect mating probability than distance does (Chenault *et al.*, 2008 on *Populus nigra*). This drawback could be partially solved by improving the Bayesian approach developed here to include a supplementary unobserved random variable associated with each individual that would model the flowering time (and ideally one variable for the duration, to account both for timing and length of the flowering period). The information from typical genotypic data is expected to be sufficient to estimate preferential mating between some individuals in addition to the spatial component and the differential fecundities among pollen donors that we presently estimate (even if preferential mating can be the result of mechanisms other than phenology). Estimating this unobserved phenological data for all trees would necessitate sampling a reasonably large number of mother plants with various phenologies, probably a large number of seeds, and would require some prior information about phenological variance and overlap.

Our Bayesian approach could probably be improved by providing more concentrated prior distributions for the parameters. Here in particular, we used distributions over wide areas [(0·50 km) for the mean dispersal distance, (0·5) for the shape parameter *b*, (1, 1000) for the ratio d_{obs}/d_{ep}) with long tails providing non-negligible weight to large unrealistic values. These prior distributions could explain the bias for the dispersal parameters in high-density populations. Even if this type of prior distributions can be justified if there is a complete absence of expectation concerning the scale of pollen dispersal, in practice we generally have some preliminary knowledge about the biology of the species that could be used to provide more concentrated prior distributions. Adding the possibility to use several prior distributions in the MEMM program is a necessary improvement.

One main result of our simulations is the unexpected Type I error rate of the likelihood ratio tests (or equivalently AIC-based model selection) used in the classical mating models [e.g Chybicki & Burczyk 2010; Goto *et al.* 2006; Oddou-Muratorio, Klein & Austerlitz 2005; Shimatani *et al.* 2007]. This result was qualitatively expected: a large variance in fecundity not considered in the model implies that some trees have large mating probabilities for (some) mother plants and the numerous progenies they generate (on these mothers) appear as correlated mating. Thus, the hypothesis of independent fecundation events used to compute the likelihood is no longer true. This results in over-dispersion (McCullagh & Nelder 1989) and incorrect inferences (too narrow confidence intervals, underestimated *P*-values). In Oddou-Muratorio, Klein & Austerlitz (2005) we already discussed a possible over-dispersion and the consequences for inferences. However, we did not expect that this phenomenon could be so strong in practice. Here all but one simulation with $d_{obs}/d_{ep} > 2$ concluded wrongly that there were significant effects of the covariates considered. Such a ratio of $d_{obs}/d_{ep} > 2$ is likely to occur for numerous studies [N/N_e is reported to be generally between 2 and 10 by Frankham (1995); Shimatani (2010) also reports a large variance of fecundity not explained by DBH]. Additional simulations covering a range of scenarios showed that the type I error rate found here (i) was weakly sensitive to the dispersal kernel used, (ii) was not sensitive to the consideration of another covariate (related or unrelated to fecundity), but (iii) was lower for 'simpler' covariates (one quantitative covariate < three quantitative covariates < one class covariate < two class covariates), (Data S2).

Hereford, Hansen & Houle (2004) previously listed several reasons why selection gradients found in the literature could be generally overestimated. Here, we showed that for studies typical of trees species, the covariates reported so far as significant could also be over-represented. Thus, conclusions about significant covariates affecting fecundity should be carefully considered. We can expect that the biological relevance of the variation in fecundity detected across covariates values helps to draw wise conclusions.

Possible alternatives are proposed by statistical theory to account for these correlated matings because of variance in fecundity. A first solution can be bootstrap consisting in resampling mother plants and using the genotypes of the actual seeds from these mother plants until reaching the same number of seeds as the observed data set (e.g. Chenault *et al.* 2008). This procedure generates bootstrap data sets that keep the correlation structure among seeds within progenies. A second solution would be to develop mixed-effects mating models that simultaneously consider fixed effects of the covariates on fecundity and an additional random individual effect that accounts for the remaining unexplained part of variance in fecundity. The Bayesian framework developed here makes it possible to estimate the parameters in this statistical model, and the MEMM computer program should integrate fixed effects soon. This type of approach has already been fruitful in several domains which simultaneously consider genotypic and demographic information such as survival analysis, life-history traits estimation or heritability in relation to capture-mark-recapture data (Gimenez & Choquet 2010).

Finally, this study stresses the difference between individual fertility (i.e. male reproductive success) and individual fecundity. The former is defined as the actual number of ovules that a given plant fertilizes after pollen dispersal (either as an expected number before sampling or as a realized number after sampling, which is estimated from categorical assignments), whereas the latter is the amount of efficient pollen released before dispersal. Achieving a paternity assignment and counting the number of seeds fertilized by each pollen donor provides a good estimate of individual fertilities but is not perfectly correlated to individual fecundity because of (i) the spatial arrangement of individuals and (ii) the sampling strategy of the mother trees. Here, we found correlations between fertilities and fecundities between 0·8 and 0·9. But the differences between fecundity and expected fertility should increase with more clustered distributions of trees: The fertility of an isolated individual is lower than expected from its fecundity for instance. And the differences should increase with a decreasing number of sampled mother trees because no correction is applied to fertility estimates to account for the biases generated by a non-representative sampling (i.e. some mothers instead of all plants providing seeds). Furthermore, the differences between expected fertility and observed fertility should be more variable with a smaller number of seeds analysed. Both phenomena should amplify the poor performance of basic paternity analysis in estimating individual fecundities when the sampling rate decreases.

Because fertility is not always the characteristics of principal interest, it can be fruitful to separate the spatial (and phenological) and fecundity components within a modelling framework. In this way, it is possible to compute in a second step the male reproductive success (and the variance of male reproductive success, determining the effective size of male population) in different spatial designs, or for different sampling schemes (for instance, over all progeny of all trees in the study site). If the main goal of a study is to characterize the effective number of pollen donors per mother tree in the particular spatial configuration of the study site basic paternity analyses provide a satisfactory answer. The former approach is a first step towards a 'mechanistic model' of mating patterns, while the second is a more descriptive approach.

Acknowledgements

We are grateful to Tonya Lander for comments on the successive versions of the manuscript. This work was supported by the programme 'ECOlogie pour la Gestion des Ecosystèmes et de leur Ressources' funded by INRA.

References

Adams, W.T. & Birkes, D.S. (1991) Estimating mating patterns in forest tree populations. *Biochemichal Markers in the Population Genetics of Forest Trees* (eds S. Fineschi, M.E. Malvolti, F. Cannata & H.H. Hattemer), pp. 157–172. SPB Academic Publishing, The Hague.

Adams, W.T., Griffin, A.R. & Moran, G.F. (1992) Using paternity analysis to measure effective pollen dispersal in plant populations. *American Naturalist*, **140**, 762–780.

Austerlitz, F. & Smouse, P.E. (2002) Two-generation analysis of pollen flow across a landscape. IV. Estimating the dispersal parameter. *Genetics*, **161**, 355–363.

Austerlitz, F., Dick, C.W., Dutech, C., Klein, E.K., Oddou-Muratorio, S., Smouse, P.E. & Sork, V.L. (2004) Using genetic markers to estimate the pollen dispersal curve. *Molecular Ecology*, **13**, 937–954.

Bacles, C.F.E., Burczyk, J., Lowe, A.J. & Ennos, R.A. (2005) Historical and contemporary mating patterns in remnant populations of the forest tree *Fraxinus excelsior* L. *Evolution*, **59**, 979–990.

Broquet, T. & Petit, E.J. (2009) Molecular estimation of dispersal for ecology and population genetics. *Annual Review of Ecology Evolution and Systematics*, **40**, 193–216.

Burczyk, J., Lewandowski, A. & Chalupka, W. (2004) Local pollen dispersal and distant gene flow in Norway spruce (*Picea abies* [L.] Karst.). *Forest Ecology and Management*, **197**, 39–48.

Burczyk, J., Adams, W.T., Moran, G.F. & Griffin, A.R. (2002) Complex patterns of mating revealed in a *Eucalyptus regnans* seed orchard using allozyme markers and the neighbourhood model. *Molecular Ecology*, **11**, 2379–2391.

Chakraborty, R., Meagher, T.R. & Smouse, P.E. (1988) Parentage analysis with genetic markers in natural populations. I. The expected proportion of offspring with unambiguous paternity. *Genetics*, **118**, 527–536.

Chenault, N., Dowkiw, A., Jorge, V., Villar, M., Juteau, M., Guerin, V. *et al.* (2008) Pollen flow between Lombardy poplar and natural populations of black poplar. In *Proceedings of the IUFRO-CTIA 2008 Joint Conference*, Quebec, August 25–28, 2008.

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

Devlin, B. & Ellstrand, N.C. (1990) Male and female fertility variation in wild radish, a hermaphrodite. *American Naturalist*, **136**, 87–107.

Devlin, B., Roeder, K. & Ellstrand, N.C. (1988) Fractional paternity assignment: theoretical development and comparison to other methods. *Theoretical & Applied Genetics*, **76**, 369–380.

Ellstrand, N.C. (1992) Gene flow by pollen: implications for plant conservation genetics. *Oikos*, **63**, 77–86.

Fernandez, J. & Gonzalez-Martinez, S.C. (2009) Allocating individuals to avoid inbreeding in ex situ conservation plantations: so far, so good. *Conservation Genetics*, **10**, 45–57.

Frankham, R. (1995) Effective population-size adult-population size ratios in wildlife – A review. *Genetical Research*, **66**, 95–107.

Garcia, C., Arroyo, J.M., Godoy, J.A. & Jordano, P. (2005) Mating patterns, pollen dispersal, and the ecological maternal neighbourhood in a *Prunus mahaleb* L. population. *Molecular Ecology*, **14**, 1821–1830.

Gérard, P., Klein, E.K., Austerlitz, F., Fernandez-Majarres, J.F. & Frascaria-Lacoste, N. (2006) Assortative mating and differential male mating success in an ash hybrid zone population. *BMC Evolutionary Biology*, **6**, 96.

Gimenez, O. & Choquet, R. (2010) Individual heterogeneity in studies on marked animals using numerical integration: capture-recapture mixed models. *Ecology*, **91**, 148–154.

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.

Hereford, J., Hansen, T.F. & Houle, D. (2004) Comparing strengths of directional selection: how strong is strong? *Evolution*, **58**, 2133–2143.

Johnson, J.B. & Omland, K.S. (2004) Model selection in ecology and evolution. *Trends in Ecology and Evolution*, **19**, 101–108.

Kang, K.S., Bila, A.D., Harju, A.M. & Lindgren, D. (2003) Estimation of fertility variation in forest tree populations. *Forestry*, **76**, 329–344.

Klein, E.K., Desassis, N. & Oddou-Muratorio, S. (2008) Pollen flow in the wildservice tree, *Sorbus torminalis* (L.) Crantz. IV. Whole interindividual variance of male fecundity estimated jointly with the dispersal kernel. *Molecular Ecology*, **17**, 3323–3336.

Klein, E.K., Lavigne, C. & Gouyon, P.H. (2006) Mixing of propagules from discrete sources at long distance: comparing a dispersal tail to an exponential. *BMC Ecology*, **6**, 3.

Marshall, T.C., Slate, J., Kruuk, L.E.B. & Pemberton, J.M. (1998) Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular Ecology*, **7**, 639–655.

McCullagh, P. & Nelder, J.A. (1989) *Generalized Linear Models*. Chapman & Hall/CRC, Boca Raton.

Meagher, T.R. (1986) Analysis of paternity within a population of *Chamaelirium luteum*. I. Identification of the most-likely male parents. *American Naturalist*, **128**, 199–215.

Nielsen, R., Mattila, D.K., Clapham, P.J. & Palsboll, P.J. (2001) Statistical approaches to paternity analysis in natural populations and applications to the North Atlantic humpback whale. *Genetics*, **157**, 1673–1682.

Oddou-Muratorio, S., Klein, E.K. & Austerlitz, F. (2005) Pollen flow in the wildservice tree, *Sorbus torminalis* (L.) Crantz. II. Pollen dispersal and heterogeneity in mating success inferred from parent-offspring analysis. *Molecular Ecology*, **14**, 4441–4452.

Oddou-Muratorio, S., Klein, E.K. & Austerlitz, F. (2006) Real-time patterns of pollen flow in the wildservice tree, *Sorbus torminalis* III. Mating patterns and the ecological maternal neighborhood. *American Journal of Botany*, **93**, 1650–1659.

Oddou-Muratorio, S., Aligon, C., Decroocq, S., Plomion, C., Lamant, T. & Mush-Demesure, B. (2001) Microsatellite primers for *Sorbus torminalis* and related species. *Molecular Ecology Notes*, **1**, 297–299.

Oddou-Muratorio, S., Houot, M.L., Demesure-Musch, B. & Austerlitz, F. (2003) Pollen flow in the wildservice tree, *Sorbus torminalis* (L.) Crantz. I. Evaluating the paternity analysis procedure in continuous populations. *Molecular Ecology*, **12**, 3427–3439.

Oddou-Muratorio, S., Demesure-Musch, B., Pelissier, R. & Gouyon, P.H. (2004) Impacts of gene flow and logging history on the local genetic structure of a scattered tree species, *Sorbus torminalis* L. Crantz. *Molecular Ecology*, **13**, 3689–3702.

Robledo-Arnuncio, J.J. & Austerlitz, F. (2006) Pollen dispersal in spatially aggregated populations. *American Naturalist*, **168**, 500–511.

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

Robledo-Arnuncio, J.J., Austerlitz, F. & Smouse, P.E. (2007) POLDISP: a software package for indirect estimation of contemporary pollen dispersal. *Molecular Ecology Notes*, **7**, 763–766.

Robledo-Arnuncio, J.J. & Gil, L. (2005) Patterns of pollen dispersal in a small population of *Pinus sylvestris* L. revealed by total-exclusion paternity analysis. *Heredity*, **94**, 13–22.

Savolainen, O., Pyhajarvi, T. & Knurr, T. (2007) Gene flow and local adaptation in trees. *Annual Review of Ecology Evolution and Systematics*, **38**, 595–619.

Shimatani, I.K. (2010) Spatially explicit neutral models for population genetics and community ecology: extension of the Neyman-Scott clustering process. *Theoretical Population Biology*, **77**, 32–41.

Shimatani, K., Kimura, M., Kitamura, K., Suyama, Y., Isagi, Y. & Sugita, H. (2007) Determining the location of a deceased mother tree and estimating forest regeneration variables by use of microsatellites and spatial genetic models. *Population Ecology*, **49**, 317–330.

Smouse, P.E. & Meagher, T.R. (1994) Genetic-analysis of male reproductive contributions in *Chamaelirium luteum* (L) Gray (Liliaceae). *Genetics*, **136**, 313–322.

Smouse, P.E., Meagher, T.R. & Kobak, C.J. (1999) Parentage analysis in *Chamaelirium luteum* (L.) Gray (Liliaceae): why do some males have higher reproductive contributions? *Journal of Evolutionary Biology*, **12**, 1069–1077.

Smouse, P.E. & Sork, V.L. (2004) Measuring pollen flow in forest trees: an exposition of alternative approaches. *Forest Ecology And Management*, **197**, 21–38.

Sork, V.L., Nason, J., Campbell, D.R. & Fernandez, J.F. (1999) Landscape approaches to historical and contemporary gene flow in plants. *Trends in Ecology & Evolution*, **14**, 219–224.

Streiff, R., Ducousso, A., Lexer, C., Steinkellner, H., Gloessl, J. & Kremer, A. (1999) Pollen dispersal inferred from paternity analysis in a mixed oak stand of *Quercus robur* L-and *Q-petraea* (Matt.) Liebl. *Molecular Ecology*, **8**, 831–841.

Received 30 July 2010; accepted 1 December 2010
Handling Editor: Sergei L. Kosakovsky Pond

Supporting Information

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

Figure S1. Values estimated using Kindist for the dispersal parameters δ and *b*.

Figure S2. Values estimated using Kindist for the parameter measuring the variance in fecundity, d_{ep} (left) and equivalently d_{obs}/d_{ep} (right).

Data S1. Allelic frequencies used to simulate the genotypes at six microsatellite locus.

Data S2. Additional results about Type I error rates of the likelihood ratio tests when variance in fecundity is not considered. Simulations with quantitative and qualitive covariates.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.