MOLECULAR ECOLOGY

Molecular Ecology (2011) **20**, 1248–1262 doi: 10.1111/j.1365-294X.2011.05019.x

Estimating seed and pollen movement in a monoecious plant: a hierarchical Bayesian approach integrating genetic and ecological data

EMILY V. MORAN and JAMES S. CLARK

NIMBioS, 1534 White Ave, University of Tennessee, Knoxville, TN 37996-1527, USA

Abstract

The scale of seed and pollen movement in plants has a critical influence on population dynamics and interspecific interactions, as well as on their capacity to respond to environmental change through migration or local adaptation. However, dispersal can be challenging to quantify. Here, we present a Bayesian model that integrates genetic and ecological data to simultaneously estimate effective seed and pollen dispersal parameters and the parentage of sampled seedlings. This model is the first developed for monoecious plants that accounts for genotyping error and treats dispersal from within and beyond a plot in a fully consistent manner. The flexible Bayesian framework allows the incorporation of a variety of ecological variables, including individual variation in seed production, as well as multiple sources of uncertainty. We illustrate the method using data from a mixed population of red oak (*Quercus rubra*, *Q. velutina*, *Q. falcata*) in the NC piedmont. For simulated test data sets, the model successfully recovered the simulated dispersal parameters and pedigrees. Pollen dispersal in the example population was extensive, with an average father–mother distance of 178 m. Estimated seed dispersal distances at the piedmont site were substantially longer than previous estimates based on seed-trap data (average 128 m vs. 9.3 m), suggesting that, under some circumstances, oaks may be less dispersal-limited than is commonly thought, with a greater potential for range shifts in response to climate change.

Keywords: hierarchical Bayesian models, microsatellites, parentage, population ecology, *Quercus*, seed dispersal

Received 21 May 2010; revision received 16 August 2010; accepted 5 September 2010

Introduction

Seed dispersal ability has a strong influence on migration and invasion potential in plants, while the spatial scale of gene flow via both seed and pollen has important implications for population dynamics, the maintenance of genetic diversity and the effectiveness of natural selection (Kawecki 2008). Where dispersal and gene flow are limited, genetic diversity can be quickly depleted because of drift, strong selection or a combination of the two (Gillespie 2004), especially in self-incompatible species (Sork *et al.* 2002). Immigration may improve adaptive potential by increasing genetic variation (Kimbrell & Holt 2007), but local adaptation at range limits and in marginal habitats can also be inhibited when the influx of maladapted genes from the main part of the species range exceeds the rate at which they are purged by selection (Kirkpatrick & Barton 1997; Rehfeldt *et al.* 1999; Lenormand 2002; Lopez *et al.* 2007). As species have historically responded to the strong selective pressure of climate change via both migration and local adaptation (Davis & Shaw 2001), the influence of seed and pollen dispersal on these processes is of particular interest today (Holt 1990; Skelly *et al.* 2007), but unobserved dispersal processes and genotyping errors have presented challenges. Here, we introduce a flexible Bayesian approach for estimating

Correspondence: Emily V. Moran, Fax: +865 974 9461;
E-mail: emily.moran@duke.edu

seed and pollen movement, taking into account the various types of uncertainty associated with genotyping and with the dispersal process itself.

For plants, the movement of pollen and seed is the sole means of gene flow within and between populations, while seed movement alone allows for range expansion and the colonization of new sites. Probability distributions of seed and pollen movement, known as dispersal kernels, can be challenging to estimate (Clark *et al.* 2004). Parentage information is highly informative of seed and pollen dispersal, but while partial pedigrees have been obtained through long-term field observations for some animal populations, this approach is not feasible for trees, where mating and dispersal are cryptic (Pemberton 2008). For this reason, molecular markers, particularly microsatellites, are increasingly used to infer parentage, sibship or population of origin (Dow & Ashley 1996; Streiff *et al.* 1999a; Godoy & Jordano 2001; Asuka *et al.* 2005; Bacles *et al.* 2006; Hardesty *et al.* 2006; Pairon *et al.* 2006; Selkoe & Toonen 2006; Ashley 2010).

The use of molecular markers in parentage and dispersal studies presents its own challenges. Many early parentage analyses were based on excluding adults that, at a given locus, did not share an allele with the juvenile under consideration (e.g. Dow & Ashley 1996). But the probability of genotyping error or mutation is not trivial for microsatellites (Dewoody *et al.* 2006), and simple exclusion may lead to the rejection of true relationships (Jones *et al.* 2010). Several categorical or fractional allocation parentage models have been developed to take into account factors including genotyping errors and incomplete genotyping of adults that affect the likelihood of parentage (Jones *et al.* 2010). One example of this approach is the popular parentage analysis software CERVUS (Marshall *et al.* 1998). CERVUS calculates the likelihood ratio (expressed as a LOD score) for each proposed parent based on genotype, ranking parents or parent pairs according to LOD score. Individuals for which the LOD of the most likely parent is below the critical value can be assumed to have a parent outside the genotyped population. A similar categorical allocation approach, again based solely on genotype, was used by Meagher & Thompson (1987).

While genotype-only approaches to parentage assignment can be quite effective, for many plant ecologists the goal of a parentage analysis is not the pedigree itself but rather an estimate of seed and pollen dispersal kernels (Hadfield *et al.* 2006). In plants and other sessile organisms, probability of parentage often depends on distance (Levin 1981; Goto *et al.* 2006; Ashley 2010). Simulation studies have demonstrated that, when this is the case, dispersal kernels fit to mother–offspring or mother–father distances derived from a separate parentage analysis may be strongly biased, although weighting according to sampling effort can reduce this problem (Hadfield *et al.* 2006; Jones & Muller-Landau 2008).

There has been an increasing interest in developing ‘full probability’ models that simultaneously estimate parentage and population-level parameters (including seed and pollen dispersal), as it has been demonstrated that such an approach can significantly reduce bias in both (Adams *et al.* 1992; Hadfield *et al.* 2006; Jones *et al.* 2010). One such model developed to investigate biparental gene flow in plants is the ‘seedling neighbourhood model’ of Burczyk *et al.* (2006). This model estimates an immigration rate for seed (m_s) and pollen (m_p) into neighbourhoods of a given size around seedlings and mother trees, respectively. Genotypes are assumed to be observed without error (Burczyk *et al.* 2006; Chybicki & Burczyk 2010). The original model did not characterize the full dispersal kernel, which was instead calculated using a LOD approach, as distance affected probability of parentage only within a neighbourhood (Gonzalez-Martinez *et al.* 2006). Recent modifications of the model allow pollen (or seed) immigration rates to vary between mothers (or offspring) and have enabled the dispersal kernel to be smoothly extended outside the neighbourhood (Goto *et al.* 2006). In this framework, the choice of neighbourhood size can have important consequences. If two or more potential parents (based on genotype) exist within the seedling neighbourhood, the probability that tree *i* is the mother depends on weighting factors including distance, whereas if only one potential parent exists within the neighbourhood of a seedling, it is assumed to be the mother (Burczyk *et al.* 2006). Genotyped adults outside the neighbourhood are not explicitly considered as parents (Chybicki & Burczyk 2010), which makes construction of a pedigree somewhat problematic. However, if the neighbourhood is taken to be the size of the entire plot (Oddou-Muratorio & Klein 2008), this is less of a concern. The modifications previously mentioned make it possible to estimate the full dispersal kernel using the neighbourhood model, but the authors note that differences in neighbourhood size can still make between-site comparisons challenging (Chybicki & Burczyk 2010).

In their 2006 article, Hadfield *et al.* outlined a Bayesian approach to full probability modelling, showing how data such as social status or territory location could be incorporated to simultaneously estimate parentage and population-level parameters in birds. They confirmed that joint estimation improves pedigree estimates and decreases bias in population parameters when the assumptions of the specific model are met (Hadfield *et al.* 2006). A hierarchical Bayesian approach presents a number of advantages for the study of complex processes such as dispersal, including the capacity to accommodate multiple data types and multiple sources of uncertainty with relative ease within a fully consistent framework (Clark 2005). The selection of prior distributions allows the user to make use of existing information more fully (Jones *et al.* 2010). In addition, hierarchical Bayesian models allow a smooth propagation of uncertainty, so that the breadth of the posterior distribution for a dispersal parameter reflects uncertainty in data and in parentage assignments (Clark & Gelfand 2006; Cressie *et al.* 2009)—one of the main goals of full probability modelling (Jones *et al.* 2010).

The model presented here is the first developed for plants that simultaneously estimates parentage and dispersal kernels for seed and pollen within a Bayesian framework, taking into account genotyping error and variation in individual fecundity. The model treats dispersal coherently, the same process governing seed and pollen movement both inside and outside the mapped stand. All adults are considered as potential mothers and fathers of all seedlings. As in Hadfield *et al.* (2006), genotypes are subject to both allelic dropout and mistyping error. However, we have modified the treatment of mistyping error to reflect the fact that mistyping is more likely to occur between alleles of similar length (Garant *et al.* 2001; Bonin *et al.* 2004).

We demonstrate this approach using data from a mixed-species population of red oaks (*Quercus rubra, Q. velutina, Q. falcata*) located in central North Carolina. As in similar analyses focusing on seedlings rather than seeds (Gonzalez-Martinez *et al.* 2006; Goto *et al.* 2006; Chybicki & Burczyk 2010), the estimated dispersal kernels reflect effective dispersal distances after germination and early seedling mortality. Seed-trap data do not adequately capture the seed shadow of oaks and other nut-bearing trees, as they are primarily dispersed by animals that bury seeds in shallow caches (Vander Wall 2001). Information about effective dispersal is valuable in understanding the role of dispersal by animals in this system. Moreover, current evidence for long-distance gene flow via seed and pollen in oak is conflicting, although it is generally agreed that the former is much more restricted than the latter (Ducousso *et al.* 1993; Dow & Ashley 1996; Johnson *et al.* 1997; Knapp *et al.* 2001; Streiff *et al.* 2002; Sork *et al.* 2002; Li & Zhang 2003; Nakanishi *et al.* 2004; Garcia & Houle 2005; Fernandez-Manjarres *et al.* 2006; Moore *et al.* 2007; Purves *et al.* 2007; Chybicki & Burczyk 2010). While the model presented here was developed for a self-incompatible monoecious tree, the same framework is applicable to dioecious or selfing species with minor modifications.

Materials and methods

The focal species

Red oaks (*Quercus*, section *Lobatae*) are important both as timber trees and as providers of hard mast for wildlife (Little 1980; McShea *et al.* 2006). Because oaks have large, heavy seeds, they are often regarded as being more dispersal-limited than species dispersed by wind or frugivorous birds (Sork 1984; Clark *et al.* 2004; Garcia & Houle 2005). If this is the case, spatially restricted seed dispersal could contribute to the recruitment failures (attributed primarily to changes in disturbance frequencies and increased deer herbivory) that have been observed for red oaks in many parts of their range (Abrams 1992; Elliott *et al.* 1999; McDonald *et al.* 2002; Spetich 2004; Aldrich *et al.* 2005) and may also reduce their ability to respond to climate change via range shifts (Clark *et al.* 1998a; Davis & Shaw 2001). On the other hand, while rodents usually move acorns <100 m, both blue jays (*Cyanocitta cristata*) and European jays (*Garrulus glandarius*) have been observed to cache acorns hundreds of metres to kilometres away from the mother tree (Johnson *et al.* 1997; Johnson & Webb 1989; Vander Wall 2001; Gomez 2003; Purves *et al.* 2007). The blue jay and the grey squirrel (*Sciurus carolinensis*), which both bury acorns in shallow caches, are the most important acorn dispersers in the oak-hickory forest of the Southeastern US (VanderWall 2001).

In the case of pollen dispersal, while wind-dispersed pollen can travel very long distances, realized gene flow via pollen tends to be on a much smaller scale (Ducousso *et al.* 1993; Fernandez-Manjarres *et al.* 2006). In oaks, a high percentage of seeds and juveniles is usually found to be the product of pollen movement from outside of focal stands (Dow & Ashley 1996; Streiff *et al.* 1999a; Nakanishi *et al.* 2004), but some fragmented populations have been shown to be pollen-limited (Knapp *et al.* 2001; Sork *et al.* 2002). Parentage and dispersal studies have generally shown very low selfing rates in oaks (Fernandez & Sork 2007; Chybicki & Burczyk 2010), and previous studies of *Quercus rubra* have indicated complete outcrossing (Schwarzmann & Gerhold 1991; Sork *et al.* 1993). We therefore assume, in the following analysis, that no tree can be both mother and father to a seedling.

Oak species have a high ability to hybridize within sections of the genus (Burger 1975; Cottam *et al.* 1982; Ducousso *et al.* 1993; Aldrich *et al.* 2003b; Dodd & Afzal-Rafii 2004). For this reason, we included in this study not only northern red oak (*Q. rubra*), which is abundant in the study site and has published microsatellite primers (Aldrich *et al.* 2002, 2003a), but also the two species present at the study site that are most likely to hybridize with it: black oak (*Q. velutina*) and southern red oak (*Q. falcata*). Genetic structure analyses for oak species at Duke Forest show almost no between-species differentiation in allele frequencies at the six microsatellite loci under consideration, supporting the hypothesis that the three species hybridize at this site (Moran *et al.* in review). Previous studies have also found evidence of substantial levels of hybridization between co-occurring red oaks (Aldrich *et al.* 2003b; Dodd & Afzal-Rafii 2004). Consequently, all three species are considered as members of one interbreeding population in the analysis that follows.

The study population

The study population is in a second-growth forest established on former agricultural land in the North Carolina Piedmont, located in the Blackwood division of the Duke forest (35°58′N, 79°5′W). The tree community today consists of mature loblolly pines (*Pinus taeda*) intermixed with *Quercus*, *Acer* and other hardwoods. The stand was mapped for prior forest dynamics studies (Clark *et al.* 2004; Ibanez *et al.* 2007). For the purpose of this study, an additional 40- to 60-m border area was surveyed for oaks, regularizing the borders of the mapped stand (which was originally nonrectangular) and increasing total area to 12 ha. All trees >2 m tall have been tagged and measured, and long-term demographic data were available for all trees within the original stand area.

Sampled seedlings are located in permanent census plots. The original plot contained 124 such plots 2 m^2 in area, arrayed in cross-shaped transects crossing both gap and closed-canopy areas. Because the understory at this site is sparse, 79 additional 1-m^2 and 70 additional 7-m^2 census plots were added to increase sample size and to provide better representation of short- and long-range dispersal events. Plots were censused each spring to identify newly emerged or dead individuals.

Genetic data

Leaf tissue collected from adult trees (n = 118) and from sampled seedlings (n = 219) was stored at −80 °C prior to DNA extraction. Total genomic DNA was extracted from leaf tissue using a modified CTAB protocol (Data S1, Supporting information). Six nuclear microsatellites isolated by Aldrich *et al.* (2002, 2003a) were analysed using GeneMarker (Softgenetics). All loci were highly polymorphic, and all individuals had unique genotypes. Genotyping error rates for each locus were estimated by regenotyping many individuals. Treatment of genotyping error is further discussed below.

Model development

Genotypes and dispersal

Consider a population in which mature individuals I produce both pollen and seeds. These adult trees exist in a mapped area that is exhaustively sampled (all adults genotyped). Adult trees are characterized by genotype and location $\{(G_{i,l}, s_i), i = 1, \ldots, n; l = 1, \ldots L\}$, where $\mathrm{s}_i = (x_i, y_i)$ are map coordinates, l are loci, $\mathrm{G}_{i,l} = (a_{1i}, a_{2i})_l$ is the length two vector of alleles at locus l, and $(a_{1il}, a_{2il}) \in A_l$, and A_l is the set of all n_l alleles in the population at that locus. The frequency of alleles in the population at locus l is the length n_l vector $\text{freq}(a_l) = [a_{l1}, \ldots, a_{ln_l}]$, each element being equivalent to the probability of drawing allele $1\ldots n_l$ at random from the population. Assuming alleles are independent (as one would expect in an outbreeding population), the probability of a given genotype (a_1,a_2), drawn at random from the population, will be $p(G_l) = \text{freq}(a_{1l})\text{freq}(a_{2l})$. In addition to the adult trees, there is a sample of seedlings k = 1, …, K, each characterized not only by genotype G_k and location s_k, but also by pedigree, where $P_k = (i', i)$ indicates that k has mother i and father i'. The pedigree is not known, but rather will be estimated based on genotype and dispersal. The genotype of k at a given locus consists of one allele contributed by each parent.

Any adult individual $i \subseteq I$ can serve as a mother or a father. Pollen released from individual i' may disperse to and fertilize a flower from individual i. Because of self-incompatibility, $i' \neq i$ in this example, but this assumption of exogamous pollen could be relaxed in selfing species by allowing that $i' = i$ with some probability q and that $i' \neq i$ with probability $1-q$. We assume that the probability of fertilization of individual i by i' depends on dispersal distance $d_{i,i'} = \|s_{i'} - s_i\|$. The probability that seeds are dispersed from mother i to the location of offspring k depends on distance $d_{ik} = \|s_k - s_i\|$. Other physical factors, such as height or wind direction, may be relevant in some situations, and dispersal functions can be constructed that take these into account (Cousens *et al.* 2008). However, for the sake of simplicity, we focus here on distance and fecundity.

The seed shadow for a population is equal to the sum, over all adult trees, of the number of seeds produced times the dispersal kernel expressed as probability per m^2 (Clark *et al.* 1999). Thus, the proportion of seeds expected to reach location k from tree i or the proportion of pollen received by tree i originating from tree i' depends not only on distance but on the fecundity f_i or the pollen production $c_{i'}$. Seed production, $f_{i,t}$, by all trees for years t = 2000, …, 2008 in the plot has been estimated using a separate hierarchical Bayesian model in which fecundity and probability of sexual maturity are informed by seed-trap data, observations of flowering, and tree size and growth (Clark *et al.* 2004, 2010). That model includes the 'summed seed shadow' inverse modelling approach described in Clark *et al.* (1998b, 1999). Because many of the sampled seedlings recruited before the beginning of the present study and their exact age is not known, seedling parentage effectively integrates over multiple years of seed production. We therefore incorporate variation and uncertainty in fecundity by defining a mean and standard deviation for f_i over the 2000–2008 period and, at each iteration of the Gibbs sampler, drawing a new value for f_i from this distribution (see 'Implementation' and Data S2, Supporting information). Trees that are large and fecund tend to produce more pollen than trees that are small or immature. However, detailed studies on male and female allocation within individual oaks (as opposed to at the stand level) are lacking in the literature. In the absence of more information, pollen grains produced per father per year $c_{i'}$ are assumed to be proportional to estimated seed production, $f_{i'}$, for the same individual. Genotype data are the ultimate arbiter of whether a tree that is known to be reproductively mature is a potential mother or father for a given seedling.

We now consider the probability of pedigree P_k, i.e. the probability that i is the mother and i' is the father of k, which depends on the genotypes of all three individuals weighted by any other factors that affect the probability that individual k could have parents (i',i). In this example, the probability that a seedling has parent pair (i',i), before we know anything about genotype, is taken to depend on seed and pollen production of the proposed parents and the probability of pollen movement over distance $d_{i'i}$ and of seed movement over distance d_{ik}:

$$p(d_{i'i,}d_{ik}|u_\mathrm{s}, u_\mathrm{p}, P_k) = \frac{c_{i'}p\left(d_{i'i}|u_\mathrm{p}\right)f_i p(d_{ik}|u_\mathrm{s})}{\sum\limits_{i'\in I}\sum\limits_{i\in I} c_{i'}p\left(d_{i'i}|u_\mathrm{p}\right)f_i p(d_{ik}|u_\mathrm{s})} \qquad (1)$$

where u_p is the pollen dispersal parameter and u_s is the seed dispersal parameter, but other criteria could be used (Hadfield *et al.* 2006). This probability is expressed as a ratio, relative to all other potential parents.

Given that i and i' are parents of k and that individuals are genotyped at L loci, the probability of the offspring genotype given the pedigree is:

$$p(G_k|P_k = (i',i), G_{i'}, G_i) \propto \prod_{l=1}^{L} p(G_{kl}|P_k = (i',i), G_{i'l}, G_{il}) \qquad (2)$$

The two sides of eqn 2 are expressed as a proportionality, because the probability will be normalized over all potential parent pairs. The factors on the right-hand side of eqn 2 are the standard Mendelian probabilities for diploid organisms. Note that we could swap subscripts i and i', representing the equivalent case for the mother being the father and vice versa, and the probability of producing a given offspring genotype would be the same. Probabilities are not equivalent once dispersal is considered, because dispersal probabilities of seed and pollen differ. Given that i and i' could have produced an offspring of genotype G_k, the likelihood that this pair is the true parents relative to all other possible parent pairs depend on the dispersal kernels for seed and pollen and the seed and pollen production of all trees.

The dispersal kernel is a density function, representing the probability (per m^2) of seed or pollen travelling a given distance from the parent tree. Previous studies show that for animal dispersed seed and wind-dispersed pollen the dispersal kernel is usually fat-tailed, with both more short-distance and more long-distance events than in a Gaussian distribution (Clark *et al.* 1999; Goto *et al.* 2006; Hardesty *et al.* 2006; Streiff *et al.* 1999a). For this reason, and to facilitate comparison with previous work by Clark *et al.* (1999, 2001, 2005), a 2D-t kernel was chosen to represent both seed and pollen dispersal probabilities. The probability of pollen or seed travelling a given distance d is given as:

$$p(d) = \frac{1}{\pi u\left(1 + \frac{d^2}{u}\right)^2} \qquad (3)$$

where the shape of the kernel is determined by the parameter u (u_p for pollen, u_s for seed). The mean dispersal distance is given by:

$$E(d) = \pi/2\sqrt{u} \qquad (4)$$

The general model structure can accommodate other types of distributions, such as Gaussian or power-exponential, but we do not address these here. More information about the 2D-t kernel can be found in Clark *et al.* (1999), while Cousens *et al.* (2008) provide a good overview of different types of dispersal kernel.

The expected density of seed or pollen reaching a given point k from a particular source tree i is equal to the probability of the seed or pollen grain travelling the distance d_{ki}, given by the dispersal kernel, times the fecundity or pollen production of the source tree. Because the expected amount of seed is in units of seeds/m^2, this quantity is multiplied by the size of the plot to approximate the number of seeds expected to reach that plot.

Because focal populations in population-genetic studies are seldom completely isolated, it is important to allow for the possibility that parents of a sampled offspring reside outside the sampled area. In this model, we assume that the sampled area is part of a continuous population and that the density of adult trees outside the plot is equal to the density of adults inside, allowing us to approximate expected seed and pollen received from out-of-plot sources via numerical integration (Data S3, Supporting information). This assumption is appropriate when dealing with continuous forest, as in the present example, but may not be justified in all situations. If information exists about the distribution of out-of-plot seed or pollen sources, this can and should be included.

Genotype error

Genotype errors in microsatellites (Fig. 1) are predominantly of two varieties: mistyping causes one allele to be mistaken for another (usually of similar length), while allelic dropout causes a heterozygote to look like a homozygote (Dewoody *et al.* 2006). Both error rates can be estimated by repeated genotyping of individuals and loci (Bonin *et al.* 2004). This was carried out for all six loci, using data from two study populations in North Carolina (Moran & Clark in review). Across loci, mistyping occurred at an average of 5.7% (range 2–18%) of regenotyped alleles and dropout at 5% (range 2–8%) (Table S1.2 in Data S1, Supporting information). These rates are high, but microsatellites often exhibit high error rates (Bonin *et al.* 2004; Burczyk *et al.* 2004; Dewoody *et al.* 2006). In this case, the high concentrations of tannins and other secondary compounds in oak leaves made it challenging to obtain clean DNA samples of consistent concentration, and amplification success for a single individual could vary considerably from one extraction to another (see Data S4, Supporting information). We develop models for both main error types.

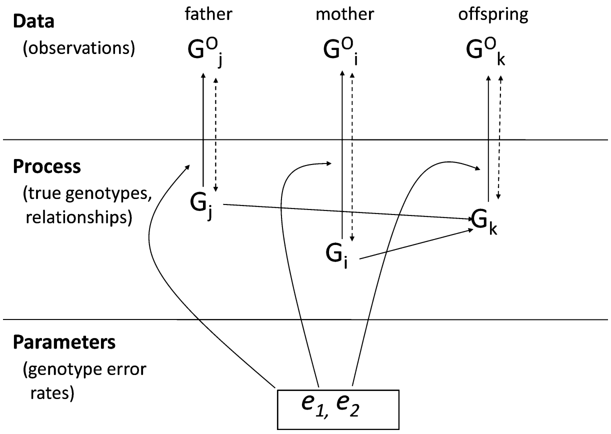

Fig. 1 Relationship between true and observed genotypes. Dashed arrows indicate that we can calculate the probability of a given true genotype given the observed genotype, as well as the probability of observing a certain genotype given the truth.

Mistyping occurs when an allele is amplified using PCR and some copies are longer or shorter than the true length. This ‘stutter’ can cause the length of an allele to be misread by one repeat length (Garant *et al.* 2001)—in this case, two base pairs. Previous models (Marshall *et al.* 1998; Hadfield *et al.* 2006) have generally assumed that if an allele is mistyped, the probability of observing any ‘false’ allele is proportional to the frequency of that allele in the population. However, because it is unlikely that the observed allele will differ greatly in length from the true allele except in the rare case of contamination or sample mislabelling (Garant *et al.* 2001; Bonin *et al.* 2004), we assume that only alleles adjacent in length (differing by 1–2 bp and expressed in Table 1 as $a^o = a \pm 1$) can be mistaken for one another. Differences of one repeat length between parent and offspring or between two samples from the same tree may also occur because of mutation. Microsatellite markers have high mutation rates, which generate the high intrapopulation variation that makes them useful for parentage analysis (Jones *et al.* 2010). However, by allowing for a relatively high rate of mistyping, we prevent the inadvertent exclusion of potential parents because of either mistyping or mutation.

Allelic dropout occurs when one of the two alleles at a locus fails to amplify (expressed in Table 1 as $a^o = 0$). Like mistyping, this error rate can be estimated by regenotyping multiple individuals and loci, because frequently the allele that was missed on the first genotyping will be detected in the second and vice versa. The probability that a heterozygote will appear to be a homozygote in our model is based on this regenotyping data. Null alleles can also cause a heterozygote to be typed as a homozygote and are more difficult to identify because they *never* amplify (but see Chybicki *et al.* 2009). The presence of null alleles is suggested by an excess of homozygotes in a population, although this can also result from inbreeding. As with mistyping and mutation, our method of treating allelic dropout will ensure that individuals that are homozygous because of null alleles are not eliminated as potential parents or offspring, but it should be noted that the probability of a heterozygote being identified as a homozygote calculated by regenotyping may be an underestimate.

Let G_i^o be the observed genotype, which can differ from the true genotype of individual i by mistyping or dropout error. A mistyping error (event E_1) occurs with probability $p(E_1) = e_1$ and a dropout (event E_2) with probability $p(E_2) = e_2$. These probabilities are taken as fixed for each locus and are determined by regenotyping many individuals and loci. For two alleles at each locus, define the matrix $\mathbf{E} = \begin{pmatrix} E_{1,a1}, E_{2,a1} \\ E_{1,a2}, E_{2,a2} \end{pmatrix}$ of binary

Table 1 Genotyping error probabilities

	No error	1 Allele mistyped	1 Allele dropped	1 Allele dropped, the other mistyped	Both alleles mistyped
Observed genotypes	$(a_1^{\mathrm{o}} = a_1,$ $a_2^{\mathrm{o}} = a_2)$	$(a_1^{\mathrm{o}} = a_1 \pm 1, a_2^{\mathrm{o}} = a_2)$ or $(a_1^{\mathrm{o}} = a_1, a_2^{\mathrm{o}} = a_2 \pm 1)$	$(a_1^{\mathrm{o}} = a_1, a_2^{\mathrm{o}} = 0)$ or $(a_1^{\mathrm{o}} = 0, a_2^{\mathrm{o}} = a_2)$	$(a_1^{\mathrm{o}} = a_1 \pm 1, a_2^{\mathrm{o}} = 0)$ or $(a_1^{\mathrm{o}} = 0, a_2^{\mathrm{o}} = a_2 \pm 1)$	$(a_1^{\mathrm{o}} = a_1 \pm 1,$ $a_2^{\mathrm{o}} = a_2 \pm 1)$
Event $\mathbf{E} = \begin{pmatrix} E_{1,a1}, E_{2,a1} \\ E_{1,a2}, E_{2,a2} \end{pmatrix}$	$\begin{pmatrix} 0,0 \\ 0,0 \end{pmatrix}$	$\begin{pmatrix} 1,0 \\ 0,0 \end{pmatrix}$ or $\begin{pmatrix} 0,0 \\ 1,0 \end{pmatrix}$	$\begin{pmatrix} 0,0 \\ 0,1 \end{pmatrix}\begin{pmatrix} 0,0 \\ 1,1 \end{pmatrix}$ or $\begin{pmatrix} 0,1 \\ 0,0 \end{pmatrix}\begin{pmatrix} 1,1 \\ 0,0 \end{pmatrix}$	$\begin{pmatrix} 1,0 \\ 0,1 \end{pmatrix}\begin{pmatrix} 1,0 \\ 1,1 \end{pmatrix}$ or $\begin{pmatrix} 0,1 \\ 1,0 \end{pmatrix}\begin{pmatrix} 1,1 \\ 1,0 \end{pmatrix}$	$\begin{pmatrix} 1,0 \\ 1,0 \end{pmatrix}$
No. of observable combinations	1	4	2	4	4
Probability of each combination	$\frac{(1-e_1)^2(1-e_2)^2}{1-e_2^2}$	$\frac{e_1(1-e_1)(1-e_2)^2}{2(1-e_2^2)}$	$\frac{((1-e_1)^2 e_2(1-e_2))}{1-e_2^2} +$ $\frac{(e_1(1-e_1)e_2(1-e_2))}{1-e_2^2}$	$\frac{e_1(1-e_1)e_2(1-e_2)}{2(1-e_2^2)} +$ $\frac{e_1^2 e_2(1-e_2)}{2(1-e_2^2)}$	$\frac{e_1^2(1-e_2)^2}{4(1-e_2^2)}$
Total probability $p(\mathbf{E})$	$\frac{(1-e_1)^2(1-e_2)^2}{1-e_2^2}$	$\frac{2e_1(1-e_1)(1-e_2)^2}{1-e_2^2}$	$\frac{2(1-e_1)e_2(1-e_2)}{1-e_2^2}$	$\frac{2e_1 e_2(1-e_2)}{1-e_2^2}$	$\frac{e_1^2(1-e_2)^2}{1-e_2^2}$

indicators, where the first row represents one allele and the second row the other allele.

Both types of errors can occur for either allele. Both errors *can* occur simultaneously at a locus, but if two alleles are observed, then we know that event E_2 has not occurred, because dropout always results in the appearance of a homozygote even if the other allele has been mistyped. If a mistyping and a dropout event were to occur at the same allele, only the dropout event will be observed. In Table 1, we take this event into account. When neither allele at a locus is observed, this could be because of the same processes that cause only one allele to amplify, but it may also be because of other causes—a badly degraded sample, for instance. We therefore assume that we do not observe the case where both alleles drop out, which occurs with probability

$$p\begin{pmatrix} 0,1 \\ 0,1 \end{pmatrix} + p\begin{pmatrix} 1,1 \\ 0,1 \end{pmatrix} + p\begin{pmatrix} 0,1 \\ 1,1 \end{pmatrix} + p\begin{pmatrix} 1,1 \\ 1,1 \end{pmatrix} = e_2^2;$$

therefore, the probabilities in Table 1 are normalized by $(1 - e_2^2)$. Notice that the consequences of each possible event are different for homozygotes and heterozygotes.

The full model for all individuals is therefore:

$$p(P, u_{\mathrm{s}}, u_{\mathrm{p}} | \{G^{\mathrm{o}}\}, \{d\}) \propto p(\{d\} | u_{\mathrm{s}}, u_{\mathrm{p}}, P) p(\{G^{\mathrm{o}}\} | P) p(u_{\mathrm{s}}) p(u_{\mathrm{p}}) \tag{5}$$

where $\{d\}$ is the set of distances between pairs of individuals, u_{s} and u_{p} are seed and pollen dispersal parameters, respectively, $\{G^{\mathrm{o}}\}$ is the set of observed genotypes of all individuals, and $p(u_{\mathrm{s}})$ and $p(u_{\mathrm{p}})$ are Gaussian priors on the dispersal parameters. Priors were constructed based on data from the literature (Darley-Hill & Johnson 1981; Dow & Ashley 1996; Fernandez-Manjarres *et al.* 2006; Li & Zhang 2003; Moore *et al.* 2007; Nakanishi *et al.* 2004; Streiff *et al.* 1999b). We assigned u_{s} a prior mean of 253, corresponding to a mean dispersal distance of 25 m, and a prior standard deviation of 2000, truncated at 10 and 10 000. We assigned u_{p} a prior mean of 1000, corresponding to a mean dispersal distance of 70.2 m, and a prior standard deviation of 1500, truncated at 10 and 15 000 (see Data S4 for more details, Supporting information).

In expanded format, the model may be written as:

$$\begin{aligned} & p(P, u_{\mathrm{s}}, u_{\mathrm{p}} | \{G^{\mathrm{o}}\}, \{d\}, e_1, e_2, \{f\}, \{c\}) \\ & \quad \propto \prod_k \left[\left(\frac{c_{i'} p(d_{i'i} | u_{\mathrm{p}}) f_i p(d_{ik} | u_{\mathrm{s}})}{\sum_{i,i'} c_{i'} p(d_{i'i} | u_{\mathrm{p}}) f_i p(d_{ik} | u_{\mathrm{s}})} \right) \right. \\ & \qquad \left. \times \left(\frac{\prod_l p(G_{k,l}^{\mathrm{o}} \middle| G_{i',l}^{\mathrm{o}}, G_{i,l}^{\mathrm{o}}, e_{1,l}, e_{2,l})}{\sum_{i,i'} \prod_l p(G_{k,l}^{\mathrm{o}} \middle| G_{i',l}^{\mathrm{o}}, G_{i,l}^{\mathrm{o}}, e_{1,l}, e_{2,l})} \right) \right] p(u_{\mathrm{s}}) p(u_{\mathrm{p}}) \end{aligned} \tag{6}$$

Thus far, the model appears computationally demanding, because conditional probabilities are expressed in terms of latent genotypes (eqns 1 and 2), which are observed with error and therefore must be estimated. Implementation with MCMC would require substantial overhead to sample latent variables because of the large number of potential pedigree combinations. These true states do not appear in eqn 6 because we can marginalize them away, expressing observed offspring genotype conditioned directly on observed parent genotypes. Here, we demonstrate that this is the case.

Consider the factor of eqn 6 relating to the probability of the observed offspring genotype given the observed genotypes of the proposed parents. Using the observed genotypes and the genotyping error distributions, we can calculate the probability that a given pair of parents could give rise to an observed offspring genotype:

$$p(G_k^{\mathrm{o}}|G_i^{\mathrm{o}},G_j^{\mathrm{o}})=\prod_l\sum_{G_{k,l}}p(G_{k,l}^{\mathrm{o}}|G_{k,l})\sum_{G_{i,l}}\sum_{G_{j,l}}p(G_{k,l}|G_{i,l},G_{j,l}) \times p(G_{j,l}|G_{j,l}^{\mathrm{o}})p(G_{i,l}|G_{i,l}^{\mathrm{o}})=\prod_l p(G_{k,l}^{\mathrm{o}}||G_{i,l}^{\mathrm{o}},G_{j,l}^{\mathrm{o}}) \tag{7}$$

The probabilities $p(G_l^{\mathrm{o}}|G_l)$are contained in Table 1. To obtain $p(G_l|G_l^{\mathrm{o}})$, we use Bayes theorem:

$$p(G_l|G_l^{\mathrm{o}})=\frac{p(G_l^{\mathrm{o}}|G_l)P(G_l)}{\sum_G p(G_l^{\mathrm{o}}|G_l)P(G_l)} \tag{8}$$

When an individual has been genotyped more than once at a given locus, we assume that these observations are independent:

$$p(G_l\middle|G_{1,l}^{\mathrm{o}},G_{2,l}^{\mathrm{o}})=p(G_l\middle|G_{1,l}^{\mathrm{o}})p(G_l\middle|G_{2,l}^{\mathrm{o}})$$

Marginalizing away, the true states allow us to build efficient algorithms for posterior simulation.

Implementation

Computation was implemented in R. Given the number of potential parents, offspring and loci under consideration, calculation of $P(G_k^{\mathrm{o}}|G_i^{\mathrm{o}},G_{i'}^{\mathrm{o}})$is computationally expensive. As these probabilities are independent of the dispersal parameters, they were evaluated before MCMC simulation, as described later. For each offspring, we create an $(n_{\mathrm{a}}+1)\times(n_{\mathrm{a}}+1)$ matrix Amat$_k$, where n_{a} is the number of genotyped adults. Amat$_k[i,i']$ represents the probability of obtaining the observed genotype of offspring k given $\mathrm{P}_k=(i,i')$ relative to all possible parent combinations and where row $(n_{\mathrm{a}}+1)$ represents a hypothetical out-of-plot mother and column $(n_{\mathrm{a}}+1)$ a hypothetical out-of-plot father.

First, $p\left(G_{i,l}\middle|G_{i,l}^{\mathrm{o}}\right)$ is calculated at a given locus l for all potential true genotypes G_l for each adult i using eqn 8 and the probabilities given in Table 1. If an adult is ungenotyped at locus l, or if i represents a hypothetical ungenotyped out-of-plot parent, then

$$p\left(G_{i,l}\middle|G_{i,l}^{\mathrm{o}}\right)=p(G_l)=\mathrm{freq}\left(G_{1,l}\right)\mathrm{freq}\left(G_{2,l}\right).$$

Then, for the parent pair (i,i'), we calculate $p\left(G_{k,l}\middle| G_{i,l.},G_{j,l}\right)$ using Mendelian inheritance probabilities and $p\left(G_{i,l}\middle|G_{i,l}^{\mathrm{o}}\right)$ and store these probabilities in an $n_l\times n_l$ matrix, Nmat. We then calculate $p\left(G_{k,l}^{\mathrm{o}}\middle|G_{k,l}\right)$ for each offspring using the probabilities in Table 1 and store these probabilities in an $n_l\times n_l$ matrix, Omat. Finally,

$$p\left(G_{k,l}^{\mathrm{o}}\middle|G_{i,l}^{\mathrm{o}},G_{i',l}^{\mathrm{o}}\right)=\sum_G \mathrm{Nmat}[a1,a2]\mathrm{Omat}[a1,a2]$$

and Amat$_k[i,i']$ =

$$\prod_L p\left(G_{k,l}^{\mathrm{o}}\middle|G_{i,l}^{\mathrm{o}},G_{i',l}^{\mathrm{o}}\right)=p(G_k^{\mathrm{o}}|G_i^{\mathrm{o}},G_{i'}^{\mathrm{o}})$$

The MCMC was then implemented in the following sequence:

1. Initialize chain
An initial pedigree $P_k=(m_k,f_k)$ is generated for each seedling using Amat$_k$, with a random draw (m_k,f_k) ~ multinom(Amat$_k$). Then for each step in the Gibbs sampler:

2. Draw values for f_i, c_i
Distributions of fecundity values reflecting both year-to-year variation and uncertainty in annual fecundity estimates are developed for each tree as described in Data S2 (Supporting information). A new value for f_i is drawn at the beginning of each Gibbs step to mix over this variation and uncertainty; c_i is assumed to be proportional to fecundity.

3. Sampling of u_{s},u_{p} conditioned on P_k
Dispersal parameters are sampled with a metropolis step from the conditional distribution:

$$p(u_{\mathrm{s}},u_{\mathrm{p}}|P) =\prod_k\frac{c_{i'}p\left(d_{i'i}\middle|u_{\mathrm{p}}\right)f_ip\left(d_{ik}\middle|u_{\mathrm{s}}\right)}{\sum_{i'}\sum_i c_{i'}p\left(d_{i'i}\middle|u_{\mathrm{p}}\right)f_ip\left(d_{ik}\middle|u_{\mathrm{s}}\right)}p(u_{\mathrm{p}}|m_{\mathrm{p}},s_{\mathrm{p}})p(u_{\mathrm{s}}|m_{\mathrm{s}},s_{\mathrm{p}})$$

where i and i' are the currently imputed parents, m_p and m_s are the prior means, and s_p and s_s are the prior standard deviations. A Gaussian jump distribution is used to propose new values of u_{p} and u_{s}, and the conditional probabilities are compared. If $p_{\mathrm{new}}>p_{\mathrm{now}}$, where p_{new} is the conditional probability of the proposed values and p_{now} is the conditional probability of the current values, the proposed values are accepted. If $p_{\mathrm{new}}<p_{\mathrm{now}}$, the proposed parameter values are accepted with probability $a=p_{\mathrm{new}}/p_{\mathrm{now}}$.

4. Sampling of P_k conditional on u_{s},u_{p}
Each seedling has a currently imputed pedigree—a mother/father pair (i,i'). For the purposes of proposing new pedigree values, an $(n_{\mathrm{a}}+1)\times(n_{\mathrm{a}}+1)$ matrix, ppmat$_k$, is created for each seedling such that ppmat$_k$ $[x,y]=1$ if Amat$_k$ $[x,y]>0$; otherwise,

ppmat$_k$ [x,y] = 0. A new pedigree is proposed from (i*,i'*) ~ multinom(ppmat$_k$). This step speeds convergence by avoiding proposing parent pairs deemed impossible based on genotype, while allowing all combinations of parents *not* ruled out by genotype to be explored.

We then evaluate the conditional probability of the proposed pedigree relative to the current pedigree, given the currently imputed dispersal parameters using:

$$p(P_k=(i,i')|u_\mathrm{s},u_\mathrm{p})=p(d_{i'i},d_{ik}|u_\mathrm{s},u_\mathrm{p},P_k)p(G_k^\mathrm{o}|G_i^\mathrm{o},G_{i'}^\mathrm{o})$$
$$=\frac{c_{i'}p\left(d_{i'i}|u_\mathrm{p}\right)f_ip(d_{ik}|u_\mathrm{s})}{\sum\limits_{i',i}c_{i'}p\left(d_{i'i}|u_\mathrm{p}\right)f_ip(d_{ik}|u_\mathrm{s})}\,\frac{p(G_k^\mathrm{o}|G_i^\mathrm{o},G_{i'}^\mathrm{o})}{\sum\limits_{i,i}p(G_k^\mathrm{o}|G_i^\mathrm{o},G_{i'}^\mathrm{o})}$$

for father i' and mother i. The proposed values are accepted or rejected for each seedling as described in the previous step.

5. Steps 2–4 are repeated until the chains converge.

Simulation

Multiple simulations were conducted from different initial conditions to assure that chains converged to the posterior distribution (Data S5, Supporting information). Estimates of u_s and u_p converged quickly—generally within 100–2000 steps, depending on initial conditions. Testing with simulated data sets showed that the approach assigned the highest probabilities to the correct parent pair 97% of the time on average. Incorrect parentage assignment was usually caused by a large number of genotyping errors or ungenotyped loci in the parent–offspring pair (>3 mismatches or missing values). For an average of 86% of seedlings in a given simulation, the most frequently identified mother and father were the true mother and father, whereas for 11% the parents were 'inverted'—the true mother identified as the father and vice versa. Inversions occur because the only information we have that can help to identify mother vs. father is their location; however, the occurrence of inversions did not have large effects on the dispersal parameter estimates. For simulations in which the stand dimensions and plot number were those of the actual Duke Forest stand, the true u_s fell within the 95% CI of the dispersal estimate in all simulations. Estimates deteriorated as stand area declined.

Application to field data

The mapped plot contained 118 potential parent trees, while seedling plots contained 219 red oak seedlings. Multiple independent runs were performed with different initial values for parentage and dispersal parameters, to ensure model convergence. The chains were run for a total of 50 000 steps, with a burn-in of 30 000 steps.

Results

Independent runs show that both parentage and dispersal estimates converged to the posterior distributions. For 16% of the 219 genotyped seedlings, the estimated parents were both genotyped, in-plot adults. For 19.6% of seedlings, the father was estimated to be an in-plot individual and the mother an out-of-plot (unsampled) individual, while for 27.4% the father was identified as an out-of-plot individual and the mother as an in-plot individual. For 37% of seedlings, neither parent was estimated to be among the genotyped trees within the 12 ha plot. In-plot mother–offspring pairs are shown in Fig. 2. It should be noted that, for parentage, the posterior takes the form of a multinomial for each seedling. The 'estimated parents' are the parent pair with the highest posterior probability.

The posterior mean for the seed dispersal parameter, u_s, was 6300 (95% CI 5380–7220), corresponding to a mean dispersal distance of 127.7 m. The lower and upper bounds of the 95% credible interval correspond to mean dispersal distances of 115–133 m. The posterior mean for the pollen dispersal parameter, u_p, was 12 900 (95% CI 11 880–13 920), corresponding to a mean dispersal distance of 178.2 m. The lower and upper bounds of the 95% credible interval correspond to mean pollen dispersal distances of 171–185 m. Estimated dispersal kernels, with credible intervals, are shown in Fig. 3. Notice that the pollen dispersal kernel,

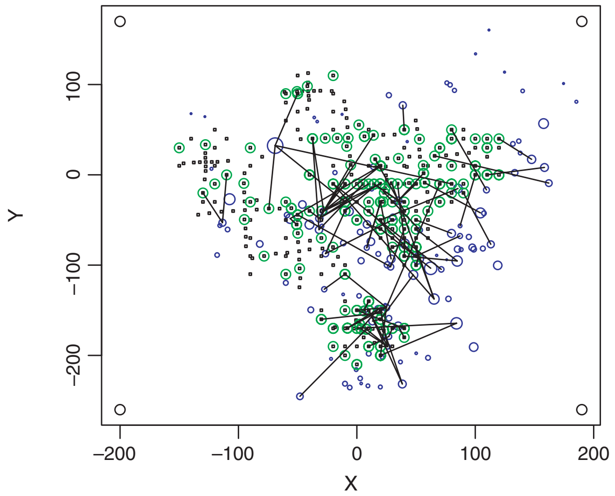

Fig. 2 In-plot mother–offspring pairs (black lines). Blue circles—adult trees. Green circles—seedlings. Black squares—seedling sampling plots. Black circles indicate corners of mapped stand.

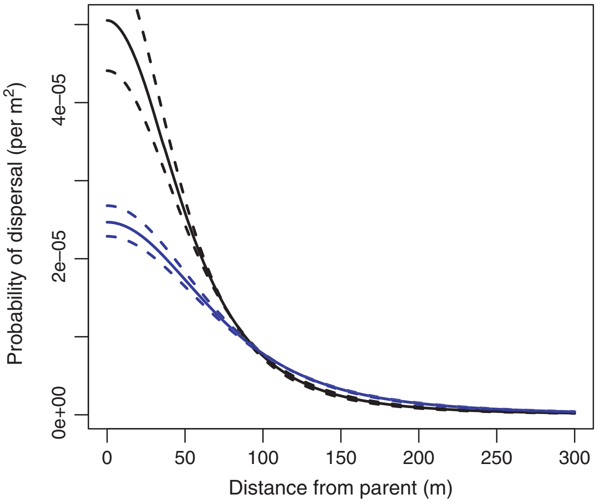

Fig. 3 Fitted 2Dt dispersal kernels for seed (black) and pollen (blue). Dashed lines—95% CI. Note that while this figure is truncated at 300 m to focus on the differences at short distances, the tails of both distributions extend much further.

in blue, is much flatter than the seed dispersal kernel at short distances, whereas at longer distances probabilities of both seed and pollen dispersal decline. However, if seed and pollen production is high, both kernels allow for a relatively high level of long-distance gene flow because of their fat tails (Clark *et al.* 2001). Posterior distributions for both dispersal parameters diverged substantially from prior distributions, indicating that data were highly informative (Data S4, Supporting information).

The mean distance between mothers and offspring within the plot was 72.4 m (range 3.1–248 m), and the mean observed father–mother distance was 101.6 m (range 8.7–229 m). These values are both shorter than the means for the dispersal kernel, as the overall estimate takes into account dispersal from outside the plot. For comparison, the average distance from a seedling to the nearest adult tree was 14.4 m (max 374 m), and the average nearest-neighbour distance between adults was 14.9 m (max 413 m).

Discussion

The Bayesian 'full probability' approach presented here combines a number of useful features not previously found within any single model of plant dispersal and parentage. It simultaneously estimates parentage and dispersal kernels for seed and pollen, making full use of both genetic and ecological data. Unlike some of the full probability models currently in use (Jones *et al.* 2010), it explicitly takes into account the two most common types of genotyping error affecting microsatellite markers, mistyping and dropout. It also takes into account the fact that mistyping errors are more likely to occur between alleles of similar length, because of PCR stutter (Garant *et al.* 2001; Bonin *et al.* 2004). The use of numerical integration (described fully in Data S3, Supporting information) enables consistent treatment of the dispersal process both inside and outside the plot, which is critical if the dispersal kernel is to reflect both long- and short-distance movement. Previous parentage studies have shown that it is often not appropriate to assume that the closest parent is the mother (Ashley 2010). Our model makes no such assumptions, and all adults are considered as potential mothers and fathers. Finally, the flexible Bayesian framework enables the inclusion of prior information about the dispersal process and a coherent treatment of uncertainty (Hadfield *et al.* 2006).

Example population: red oak at Duke Forest

The estimated seed dispersal parameter in this study ($u_s = 6300$, mean distance 124.7 m) was considerably higher than previously estimated using inverse modelling of seed-trap data ($u_s = 34.9$, mean distance 9.28 m) (Clark *et al.* 2010). It is not unexpected that genetic analyses should reveal longer effective dispersal distances, as seed-trap data for *Quercus* reflect only the initial pattern of seedfall before secondary dispersal by vertebrates, and the fit for seed-trap-derived dispersal kernels in *Quercu*s was the poorest of all taxa occurring in our North Carolina plots (Clark *et al.* 1998b). Still, the difference between the gravity-created seed shadow and effective dispersal kernel at this site is striking.

The high parent–offspring distances observed could be partly due to density-dependent mortality acting between germination and the time of sampling (Connell 1978; Janzen 1970). Distance-dependent mortality (because of adults harbouring pests or pathogens) is likely to be of minor importance to this population: seedlings exhibit 85% survival in their first year and 95.9% annual survival thereafter even though half are located within 14 m of an adult, and none are more than 60 m from an adult (unpublished data). In addition, some true in-plot parents may have died, leading to an overestimate of the number of out-of-plot parents. Many of the seedlings sampled were at least 5 years old; no mast year occurred during the 3 years of this study, and few new seedlings were produced. Because oaks lack a seed bank (Hille Ris Lambers *et al.* 2005), first-year seedlings can be assumed to have living parents, so in future studies, it would be desirable to focus on newly recruited seedlings following a mast year.

These caveats aside, the long seed dispersal distances observed at Duke Forest are not an artefact of the model, nor do they necessarily apply to all red oaks. A second mixed-species population, located in the southern Appalachians, exhibited similarly high pollen dispersal but the average seed dispersal distance was only 15 m (Moran & Clark in review). Effective dispersal distances can vary substantially between sites because of difference in the abundance or activity of dispersal vectors or in the distribution of suitable recruitment sites (Schnabel *et al.* 1998; Cousens *et al.* 2008; Terborgh *et al.* 2008; Chybicki & Burczyk 2010). While a number of studies have found restricted seed dispersal distances in oaks, as might be expected for a heavy-seeded tree dispersed by rodents (Dow & Ashley 1996; Garcia & Houle 2005; Chybicki & Burczyk 2010), others have suggested that dispersal by birds could add a significant long-distance component to the dispersal kernel (Johnson & Webb 1989; Johnson *et al.* 1997; Gomez 2003). Blue jays are common in the Blackwood Division of the Duke Forest (http://www.duke.edu/~jspippen/birds/), and previous studies have shown that jays often transport acorns >1 km and may harvest >50% of the seed crop (Darley-Hill & Johnson 1981). Grey squirrels are also abundant at Duke Forest (Moran & Clark in review), and cache pilferage and the frequent re-caching of seeds by rodents can move a seed much further than the initial cache distance (Vander Wall 2001; Roth & Vander Wall 2005).

The pollen dispersal parameter for Duke Forest converged at a value of 12 900, corresponding to a mean effective dispersal distance of 178.2 m. Wind-blown pollen can travel extremely long distances, but because oak pollen degrades relatively quickly in UV light (Schueler *et al.* 2005), and because nearby trees may produce large amounts of pollen, effective pollen dispersal may be much shorter than physical pollen transport distances (Ducousso *et al.* 1993). In some closed-canopy oak forests, short-distance matings appear to predominate (Fernandez-Manjarres *et al.* 2006), and some sparse or fragmented oak populations show evidence of pollen limitation (Knapp *et al.* 2001; Sork *et al.* 2002). Nevertheless, most previous studies in *Quercus* species have observed high out-of-plot paternity, usually in the range of 50–70% (Dow & Ashley 1996; Streiff *et al.* 1999a; Nakanishi *et al.* 2004; Chybicki & Burczyk 2010). Despite the fact that our censused plot was larger than in previous studies (generally <6 ha, vs. 12 ha in this study), we also found a similar proportion of out-of-plot paternity: 64.4%.

Because genetic structure results showed almost no differentiation between co-occurring red oak species at Duke Forest (Moran *et al.* in review), in this analysis, all individuals were treated as potential parents and offspring, regardless of morphological species classification. Just over 14% of Duke Forest seedlings were estimated to have a parent that was classed as a different species—for example, a '*Q. velutina*' mother assigned to a '*Q. rubra*' seedling. Hybridization rates have not been estimated for red oaks, but in white oaks the rates of hybridization between co-occurring species range between < 2% (e.g. Muir & Schlotterer 2005; Curtu *et al.* 2007) and > 25% (e.g. Bacilieri *et al.* 1996, Craft & Ashley 2007). Not all seedlings matched to heterospecific parents are necessarily true hybrids, but the low amount of genetic differentiation between adults classified morphologically as different species, and the steep decline in plausible in-plot parents when heterospecific individuals are excluded suggests that interspecific gene flow has been fairly common over multiple generations. This issue is discussed at length in Moran *et al.* (in review).

Model framework: benefits and caveats

In the example analysis discussed previously, we made several simplifying assumptions, which may not be appropriate for all situations. Where data on pollen production can be obtained, this information can and should be substituted for the simplistic assumption of proportional seed and pollen production. Likewise, if data contradict the assumption of similar adult density on all sides outside the plot, these data should be incorporated. For instance, if the plot is located at a forest edge, such that the only nearby adults would be to the south and east, one might choose to consider only those directions as potential seed and pollen sources. The multinomial genotyping error probabilities can also be modified to allow for mistyping errors between alleles more than one repeat length apart, although because this will increase the number of possible parent pairs this change greatly increases run times. Calculation of the probability of offspring genotypes given parental genotypes was the most computationally expensive step.

The full model given in eqn 6 can be generalized as:

$$
\begin{aligned}
&p(P,\theta_{\mathrm{p}},\theta_{\mathrm{s}},\beta \big| \{G^{\mathrm{o}}\},\{d\},x,e_1,e_2) \\
&\quad \propto \prod_k \Bigg[\left(\frac{p(d_{i'i}|\theta_{\mathrm{p}})p(d_{ik}|\theta_{\mathrm{s}})f(x,\beta)}{\sum\limits_{i,i'} p(d_{i'i}|\theta_{\mathrm{p}})p(d_{ik}|\theta_{\mathrm{s}})f(x,\beta)} \right) \\
&\qquad \times \left(\frac{\prod\limits_l p(G^{\mathrm{o}}_{k,l}|G^{\mathrm{o}}_{i',l},G^{\mathrm{o}}_{i,l},e_{1,l},e_{2,l})}{\sum\limits_{i,i'} \prod\limits_l p(G^{\mathrm{o}}_{k,l}|G^{\mathrm{o}}_{i',l},G^{\mathrm{o}}_{i,l},e_{1,l},e_{2,l})} \right) \Bigg] p(\theta)p(\beta)
\end{aligned}
$$

The dispersal kernels for seed and pollen are characterized by sets of one or more parameters θ_s and θ_p. The function f(x,β) represents the weight provided by covariates x. In our example, $f(x,\beta)$ is simply the product of f_i and $c_{i'}$, but one could also choose a set of covariates (such as diameter, age) that are related to probability of parentage according to an equation with parameters β, much as is performed in the seedling neighbourhood model (Burczyk *et al.* 2006). Genotyping error rates e_1 and e_2 could also be treated as parameters to be estimated rather than constants (Hadfield *et al.* 2006), but keep in mind that the more parameters, the greater the amount of data needed to obtain good estimates for all parameters.

In any Bayesian analysis, it is important to carefully consider the choice of priors. In this example, priors were chosen to reflect information from previous studies indicating that pollen dispersal in oaks is generally more extensive than seed dispersal, but our results were not very sensitive to changes in the prior mean. When the number of in-plot parent–offspring pairs with zero genetic mismatches is low, very wide priors can lead to an upward drift in dispersal parameter estimates; with lower genotyping error rates or larger numbers of potential parents within the plot, this becomes less important (Data S4, Supporting information). The size of the censused plot can also affect the accuracy and precision of dispersal estimates. Simulations can help to determine how large an area may be need for the system of interest (Data S5, Supporting information, see also Clark *et al.* 1998b; Cousens *et al.* 2008). Using simulated data, we found that, for u's between 50 and 1000 and a population density similar to Duke Forest, one would need a censused area greater than 300 × 300 m and more than 150 seedling census plots to obtain consistently accurate dispersal and parentage estimates. The true Duke Forest stand was 390 × 430 m and included 273 seedling census plots, and in simulations, the correct parent pair was identified 97% of the time.

The model presented here is the first to combine simultaneous estimation of dispersal and parentage for a monoecious plant with a realistic model of genotyping error. The hierarchical Bayesian framework easily accommodates multiple types of data as well as prior information, while posterior distributions for the parameters of interest incorporate uncertainty at both the data and process level. Full probability models and hierarchical Bayesian models in particular can be computationally and mathematically demanding. However, their ability to deal with the multiple factors known to affect the probability of parentage in plants in a coherent way, and to deliver better estimates of both dispersal and parentage (Jones *et al.* 2010), will only make such models increasingly useful in the future as computing and statistical resources continue to improve.

Acknowledgements

We would like to thank John Willis and his PhD students David Lowry and Young Wha Lee for their advice and assistance in the development of the molecular methods used in this study. We also thank Jing Zhang and other former Clark laboratory technicians, as well as Lisa Bukovnik and other members of the Duke DNA Sequencing Facility staff, for their assistance in data collection. Funding by the Duke Biology Department, the Association for Women in Science, and NSF is gratefully acknowledged.

References

Abrams MD (1992) Fire and the development of oak forests. *BioScience*, **42**, 346–353.

Adams WT, Griffin AR, Moran GF (1992) Using paternity analysis to measure effective pollen dispersal in plant populations. *The American Naturalist*, **140**, 762–780.

Aldrich PR, Michler CH, Sun W, Romero-Severson R (2002) Microsatellite markers for northern red oak (Fagaceae: *Quercus rubra*). *Molecular Ecology Notes*, **2**, 472–474.

Aldrich PR, Jagtap M, Michler CH, Romero-Severson R (2003a) Amplification of North American red oak microsatellite markers in european white oaks and Chinese chestnut. *Silvae Genetica*, **52**, 176–179.

Aldrich PR, Parker GR, Michler CH, Romero-Severson J (2003b) Whole-tree silvic identifications and the microsatellite genetic structure of a red oak species complex in an Indiana old-growth forest. *Canadian Journal of Forest Research*, **33**, 2228–2237.

Aldrich PR, Glaubitz JC, Parker GR, Rhodes OE, Michler CH (2005) Genetic structure inside a declining red oak community in old-growth forest. *Journal of Heredity*, **96**, 627–634.

Ashley MV (2010) Plant parentage, pollination, and dispersal: how DNA microsatellites have altered the landscape. *Critical Reviews in Plant Sciences*, **29**, 148–161.

Asuka Y, Tomaru N, Munehara Y *et al.* (2005) Half-sib family structure of *Fagus crenata* saplings in an old-growth beech-dwarf bamboo forest. *Molecular Ecology*, **14**, 2565–2575.

Bacilieri R, Ducousso A, Petit RJ, Kremer A (1996) Mating system and asymmetric hybridization in a mixed stand of European oaks. *Evolution*, **50**, 900–908.

Bacles CFE, Lowe AJ, Ennos RA (2006) Effective seed dispersal across a fragmented landscape. *Science*, **311**, 628.

Bonin A, Bellemain E, Bronken Eidesen P *et al.* (2004) How to track and assess genotyping errors in population genetics studies. *Molecular Ecology*, **13**, 3261–3273.

Burczyk J, DiFazio SP, Adams WT (2004) Gene flow in forest trees: how far do genes really travel? *Forest Genetics*, **11**, 1–14.

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

Burger WC (1975) The species concept in *Quercus*. *Taxon*, **24**, 45–50.

Chybicki IJ, Burczyk J (2010) Realized gene flow within mixed stands of *Quercus robur* L. and *Q. petraea* (Matt.) L. revealed at the stage of naturally established seedling. *Molecular Ecology*, **19**, 2137–2151.

Chybicki IJ, Trojankiewicz M, Oleksa A, Dzialuk A, Burczyk J (2009) Isolation-by-distance within naturally established populations of European beech (*Fagus sylvatica*). *Botany*, **87**, 791–798.

Clark JS (2005) Why environmental scientists are becoming Bayesians. *Ecology Letters*, **8**, 2–14.

Clark JS, Gelfand AE (2006) *Hierarchical Modelling for the Environmental Sciences*. Oxford University Press, New York.

Clark JS, Fastie C, Hurtt G *et al.* (1998a) Reid's paradox of rapid plant migration. *BioScience*, **48**, 13–24.

Clark JS, Macklin E, Wood L (1998b) Stages and spatial scales of recruitment limitation in southern Appalachian forests. *Ecological Monographs*, **68**, 213–235.

Clark JS, Silman M, Kern R, Macklin E, Lambers J HR (1999) Seed dispersal near and far: patterns across temperate and tropical forests. *Ecology*, **80**, 1475–1494.

Clark JS, Lewis M, Horvath L (2001) Invasion by extremes: population spread with variation in dispersal and reproduction. *The American Naturalist*, **157**, 537–554.

Clark JS, LaDeau S, Ibanez I (2004) Fecundity of trees and the colonization competition hypothesis. *Ecological Monographs*, **74**, 415–442.

Clark CJ, Poulsen JR, Bolker BM, Connor EF, Parker VT (2005) Comparative seed shadows of bird-, monkey-, and wind-dispersed trees. *Ecology*, **86**, 2684–2694.

Clark JS, Bell D, Chu C *et al.* (2010) High dimensional coexistence based on individual variation: a synthesis of evidence. *Ecological Monographs*, **80**, 569–608.

Connell JH (1978) Diversity in tropical rain forests and coral reefs. *Science*, **199**, 1302–1310.

Cottam WP, Tucker JM, Santamour FS (1982) *Oak Hybridization at the University of Utah*. State Arboretum of Utah, Salt Lake City.

Cousens R, Dytham C, Law R (2008) *Dispersal in Plants: A Population Perspective*. Oxford University Press, New York.

Craft KJ, Ashley MV (2007) Landscape genetic structure of bur oak (*Quercus macrocarpa*) savannas in Illinois. *Forest Ecology and Management*, **239**, 13–20.

Cressie N, Calder CA, Clark JS, Ver Hoef JM, Wikle CK (2009) Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecological Applications*, **19**, 553–570.

Curtu AL, Gailing O, Finkeldey R (2007) Evidence for hybridization and introgression within a species-rich oak (*Quercus* spp.) community. *BMC Evolutionary Biology*, **7:218**, doi: 10.1186/1471-2148-7-218.

Darley-Hill S, Johnson WC (1981) Acorn dispersal by the blue jay (*Cyanocitta cristata*). *Oecologia*, **50**, 231–232.

Davis MB, Shaw RG (2001) Range shifts and adaptive responses to quaternary climate change. *Science*, **292**, 673–679.

Dewoody J, Nason J, Hipkins VD (2006) Mitigating scoring errors in microsatellite data from wild populations. *Molecular Ecology Notes*, **6**, 951–957.

Dodd RS, Afzal-Rafii Z (2004) Selection and dispersal in a multispecies oak hybrid zone. *Evolution*, **58**, 261–269.

Dow BD, Ashley MV (1996) Microsatellite analysis of seed dispersal and parentage of saplings in bur oak, *Quercus macrocarpa*. *Molecular Ecology*, **5**, 615–627.

Ducousso A, Michaud H, Lumaret R (1993) Reproduction and gene flow in the genus *Quercus* L. *Annales des Sciences Forestales*, **50**, 91s–106s.

Elliott KJ, Hendrick RL, Major AE, Vose JM, Swank WT (1999) Vegetation dynamics after a prescribed fire in the southern Appalachians. *Forest Ecology and Management*, **114**, 199–213.

Fernandez JF, Sork VL (2007) Genetic variation in fragmented forest stands of the andean oak *Quercus humboldtii* Bonpl. (Fagaceae). *Biotropica*, **39**, 72–78.

Fernandez-Manjarres JF, Idol J, Sork VL (2006) Mating patterns of black oak *Quercus velutina* (Fagaceae) in a Missouri oak-hickory forest. *Journal of Heredity*, **97**, 451–455.

Garant D, Dodson JJ, Bernatchez L (2001) A genetic evaluation of mating system and determinants of individual reproductive success in Atlantic Salmon (*salmo salar* L.). *Journal of Heredity*, **92**, 137–145.

Garcia D, Houle G (2005) Fine-scale spatial patterns of recruitment in red oak (*Quercus rubra*): what matters most, abiotic or biotic factors? *Ecoscience*, **12**, 223–235.

Gillespie JH (2004) *Population Genetics: A Concise Guide*, 2nd edn. Johns Hopkins University Press, Baltimore.

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

Gomez JM (2003) Spatial patterns in long-distance dispersal of *Quercus ilex* acorns by jays in a heterogeneous landscape. *Ecography*, **26**, 573–584.

Gonzalez-Martinez SC, Burczyk J, Nathan R *et al.* (2006) Effective gene dispersal and female reproductive success in Mediterranean maritime pine (*Pinus pinaster* Aiton). *Molecular Ecology*, **15**, 4577–4588.

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.

Hadfield JD, Richardson DS, Burke T (2006) Towards unbiased parentage assignment: combining genetic, behavioral and spatial data in a Bayesian framework. *Molecular Ecology*, **15**, 3715–3730.

Hardesty BD, Hubbell SP, Bermingham E (2006) Genetic evidence of frequent long-distance recruitment in a vertebrate-dispersed tree. *Ecology Letters*, **9**, 516–525.

Hille Ris Lambers J, Clark JS, Lavine M (2005) Implications of seed banking for recruitment of southern Appalachian woody species. *Ecology*, **86**, 85–95.

Holt RD (1990) The microevolutionary consequences of climate change. *Trends in Ecology and Evolution*, **5**, 311–315.

Ibanez I, Clark JS, LaDeau S, Lambers JHR (2007) Exploiting temporal variability to understand tree recruitment response to climate change. *Ecological Monographs*, **77**, 163–177.

Janzen DH (1970) Herbivores and the number of tree species in tropical forests. *The American Naturalist*, **104**, 501–526.

Johnson WC, Webb TI (1989) The role of blue jays (*Cyanocitta cristata* L.) in the postglacial dispersal of fagaceous trees in eastern North America. *Journal of Biogeography*, **16**, 561–571.

Johnson WC, Adkisson CS, Crow TR, Dixon MD (1997) Nut caching by blue jays (*Cyanocitta cristata* L.): implications for tree demography. *American Midland Naturalist*, **138**, 357–370.

Jones FA, Muller-Landau HC (2008) Measuring long-distance seed dispersal in complex natural environments: an evaluation and integration of classical and genetic methods. *Journal of Ecology*, **96**, 642–652.

Jones AG, Small CM, Paczolt KA, Ratterman NL (2010) A practical guide to methods of parentage analysis. *Molecular Ecology Resources*, **10**, 6–30.

Kawecki TJ (2008) Adaptation to marginal habitats. *Annual Review of Ecology, Evolution, and Systematics*, **39**, 321–342.

Kimbrell T, Holt RD (2007) Canalization breakdown and evolution in a source-sink system. *The American Naturalist*, **169**, 370–382.

Kirkpatrick M, Barton NH (1997) Evolution of a species' range. *The American Naturalist*, **150**, 1–23.

Knapp EE, Goedde MA, Rice KJ (2001) Pollen-limited reproduction in blue oak: implications for wind pollination in fragmented populations. *Oecologia*, **128**, 48–55.

Lenormand T (2002) Gene flow and the limits to natural selection. *Trends in Ecology and Evolution*, **17**, 183–189.

Levin DA (1981) Dispersal versus gene flow in plants. *Annals of the Missouri Botanical Garden*, **68**, 233–253.

Li HJ, Zhang Z-B (2003) Effect of rodents on acorn dispersal and survival of the Liaodong oak (*Quercus liaotungensis* Koidz.). *Forest Ecology and Management*, **176**, 387–396.

Little EL (1980) *National Audubon Society Field Guide to Trees, Eastern Region*. Alfred A. Knopf, Inc., New York.

Lopez S, Rousset F, Shaw FH, Shaw RG, Ronce O (2007) Migration load in plants: role of pollen and seed dispersal in heterogeneous landscapes. *Journal of Evolutionary Biology*, **21**, 294–309.

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

McDonald RI, Peet RK, Urban DL (2002) Environmental correlates of oak decline and red maple increase in the North Carolina Piedmont. *Castanea*, **67**, 84–95.

McShea WJ, Healy WM, Devers P *et al.* (2006) Forestry matters: decline of oaks will impact wildlife in hardwood forests. *Journal of Wildlife Management*, **71**, 1717–1728.

Meagher TR, Thompson E (1987) Analysis of parentage for naturally established seedlings of *Chamaelirium luteum* (Liliaceae). *Ecology*, **68**, 803–812.

Moore JE, McEuen AB, Swihart RK, Contreras TA, Steele MA (2007) Determinants of seed removal distance by scatter-hoarding rodents in deciduous forests. *Ecology*, **88**, 2529–2540.

Moran EV, Clark JS (in review) Between-site differences in the scale of dispersal and gene flow in a forest tree. *Journal of Ecology*.

Moran EV, Willis J, Clark JS (in review) Genetic evidence for extensive hybridization in red oaks. *American Journal of Botany*.

Muir G, Schlotterer C (2005) Evidence for shared ancestral polymorphism rather than recurrent gene flow at microsatellite loci differentiating two hybridizing oaks (*Quercus* spp.). *Molecular Ecology*, **14**, 549–561.

Nakanishi A, Tomaru N, Yoshimaru H *et al.* (2004) Patterns of pollen flow and genetic differentiation among pollen pools in *Quercus salicina* in a warm temperate old-growth evergreen broad-leaved forest. *Silvae Genetica*, **53**, 258–264.

Oddou-Muratorio S, Klein EK (2008) Comparing direct vs. indirect estimates of gene flow within a population of scattered tree species. *Molecular Ecology*, **17**, 2743–2754.

Pairon M, Jonard M, Jacquemart A-L (2006) Modeling seed dispersal of black cherry, an invasive forest tree: how microsatellites may help? *Canadian Journal of Forest Research*, **36**, 1385–1394.

Pemberton JM (2008) Wild pedigrees: the way forward. *Proceedings of the Royal Society B-Biological Sciences*, **275**, 613–621.

Purves DW, Zavala MA, Ogle K, Prieto F, Rey Benayas JM (2007) Environmental heterogeneity, bird-mediated directed dispersal, and oak woodland dynamics in mediterranean Spain. *Ecological Monographs*, **77**, 77–97.

Rehfeldt GE, Ying CC, Spittlehouse DL, Hamilton DA (1999) Genetic responses to climate change in *Pinus contorta*: niche breadth, climate change, and reforestation. *Ecological Monographs*, **69**, 375–407.

Roth JK, Vander Wall SB (2005) Primary and secondary seed dispersal of bush chinquapin (Fagaceae) by scatterhoarding rodents. *Ecology*, **86**, 2428–2439.

Schnabel A, Nason J, Hamrick JL (1998) Understanding the population genetic structure of *Gleditsia triacanthos* L.: seed dispersal and variation in female reproductive success. *Molecular Ecology*, **7**, 819–832.

Schueler S, Schlunzen KH, Scholz F (2005) Viability and sunlight sensitivity of oak pollen and its implication for pollen-mediated gene flow. *Trees*, **19**, 154–161.

Schwarzmann JF, Gerhold HD (1991) Genetic structure and mating system of Northern Red Oak (*Quercus rubra* L.) in Pennsylvania. *Forest Science*, **37**, 1376–1389.

Selkoe KA, Toonen RJ (2006) Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters*, **9**, 615–629.

Skelly DK, Joseph LN, Possingham HP *et al.* (2007) Evolutionary responses to climate change. *Conservation Biology*, **21**, 1353–1355.

Sork VL (1984) Examination of seed dispersal and survival in red oak, *Quercus rubra* (Fagaceae), using metal-tagged acorns. *Ecology*, **65**, 1020–1022.

Sork VL, Huang S, Wiener E (1993) Macrogeographic and fine-scale genetic structure in a North American oak species, *Quercus rubra* L. *Annales des Sciences Forestales*, **50**, 261–270.

Sork VL, Davis FW, Smouse PE *et al.* (2002) Pollen movement in declining populations of California Valley oak, *Quercus lobata*: where have all the fathers gone? *Molecular Ecology*, **11**, 1657–1668.

Spetich MA (2004) Upland oak ecology symposium: a synthesis. Gen. Tech. Rep. SRS-73. Asheville, NC: US Department of Agriculture, Southern Research Station.

Streiff R, Ducousso A, Lexer C *et al.* (1999a) 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.

Streiff R, Labbe T, Bacilieri R *et al.* (1999b) Within-population genetic structure in *Quercus robur* L. and *Quercus petraea* (Matt.) Liebl. assessed with isozymes and microsatellites. *Molecular Ecology*, **7**, 317–328.

Streiff R, Ducousso A, Lexer C *et al.* (2002) 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.

Terborgh J, Nunez-Iturri G, Pitman NCA *et al.* (2008) Tree recruitment in an empty forest. *Ecology*, **89**, 1757–1768.

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

This study is part of PhD thesis research conducted by E.V.M. on seed dispersal, gene flow, and hybridization in red oaks, and the implications of these processes for responses to climate change. The model presented here is the result of collaboration between E.V.M. and J.S.C.

Supporting information

Additional supporting information may be found in the online version of this article.

Data S1 DNA extraction and PCR protocol.

Data S2 Fecundities.

Data S3 Out-of-plot dispersal.

Data S4 Priors.

Data S5 Simulation.

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