# Joint estimation of paternity, sibships and pollen dispersal in a snapdragon hybrid zone

**Thomas James Ellis[1,2], David Luke Field, [1,3,4], Nicholas H. Barton[1,*]**

[1] *Institute of Science and Technology Austria, 2234 Klosterneuburg, Austria*

[2] *Gregor Mendel Institute of Molecular Plant Sciences, Doktor-Bohr-Gasse 3, 1030 Vienna, Austria*

[3] *Edith Cowen University, Perth, Australia*

[4] *Applied BioSciences, Macquarie University, NSW 2109, Australia*

[*] *To whom correspondence should be addressed: nick.barton@ist.ac.at*

January 5, 2024

## 1 Abstract

The distribution of pollen dispersal distances sets the scale for plant population dynamics. A useful approach for inferring the distribution dispersal distances is to infer the distances between mates by paternity or parentage reconstruction. This is most powerful when information about multiple properties or data types are inferred in a joint analysis. We describe an approach to jointly infer paternity, sibling relationships and population parameters, with the example of the pollen dispersal kernel in a natural population of the yellow-flowered *Antirrhinum majus striatum* and the magenta-flowered *A. m. pseudomajus*. Pollen dispersal is leptokurtic, with half of mating events occurring within 30m, but with a long tail of mating events up to 747m. We also find tentative evidence that fathers tend to be to the East of mothers, indicating that there is a bias in pollen dispersal from *A. m. pseudomajus* into

*A. m. striatum*. The scale of pollen dispersal is large enough that pollinators should encounter the full range of hybrid phenotypes in the hybrid zone, and would be sufficient for any pollinator-mediated selection to influence male or female fitness.

## 2 Introduction

Knowledge of the distribution of dispersal distances organisms or gametes travel during their lifetimes aids our understanding of natural populations because it sets the scale at which populations vary (Cain, Milligan, and Strand, 2000). For example, dispersal may enhance the response to selection by increasing genetic variance, may inhibit local adaptation via swamping by maladapted alleles, or alleviate inbreeding with nearby relatives (Kremer et al., 2012). Of particular interest is both the average distance travelled and the shape of the dispersal distribution. In plants, dispersal is often characterised by leptokurtic or 'fat-tailed' distributions, where dispersal is most likely to occur over short distances, but there is a long tail of long-range dispersal events (Clark, 1998; Austerlitz et al., 2004; Bullock et al., 2017). This leptokurtosis allows much more rapid dispersal than would be suggested by the average dispersal distance alone, with long-range migrants having a disproportionate effect on the spread of adaptive alleles (Clark, 1998; Cain, Milligan, and Strand, 2000). We thus aim to accurately characterise the shape of dispersal distributions in natural populations.

A key tool for inferring dispersal is to infer the pedigree of relationships between individuals based on genetic information from parents and offspring, because this gives a direct estimate of the distances between mates (Adams, Griffin, and Moran, 1992; Cain, Milligan, and Strand, 2000; Austerlitz et al., 2004; Pemberton, 2008). Pedigree inference is most successful when as much informative data as possible can be included in a joint analysis, such as from shared alleles between siblings, or phenotype information (Neff, Repka, and Gross, 2001; Wang, 2007). Dispersal is a clear example of this. For example, one approach would be to infer a pedigree ignoring spatial information, then measure the distances between mates, assuming the pedigree to be correct. This would overestimate average dispersal, because any candidates erroneously inferred to be parents will tend to be further apart than real parents. Alternatively, one might first infer the distribution of dispersal distances using a non-genetic approach such as mark-recapture, then use this as a prior to inform pedigree inference. This would underestimate dispersal, because such methods tend to miss dispersal events over longer distances. By inferring pedigree relationships and dispersal jointly we can incorporate the information each has about the other and improve inference of both.

Several methods exist for the joint inference of parental relationships and other parameters. Various approaches have been described to jointly infer sibling relationships with the parentage or paternity of those sibships (e.g. Emery et al., 2001; Thomas and Hill, 2002; Jones et al., 2007; Wang, 2004; Anderson and Ng, 2016; Huisman, 2017). Another approach is to include data about other biological parameters that might influence mating, such as relevant phenotypes or spatial information (Neff, Repka, and Gross, 2001; Hadfield, Richardson, and Burke, 2006). This is especially appealling because it allows us to directly address biologically relevant relationships between traits of interest and mating patterns. These approaches typically

rely on an iterative algorithm such as simulated annealing or Monte-Carlo Markov chains (MCMC) to explore different pedigree structures. This is time-consuming, especially for large samples, and there is no guarantee that the full space of plausible pedigree structures can be explored. Moreover, we currently lack a framework for utilising all three sources of information - parentage, sibships and population parameters - in a single joint analysis.

We previously described a package Fractional Analysis of Paternity and Sibships (FAPS) to jointly infer of paternity and sibships (Ellis, Field, and Barton, 2018), which we here extend to include inference of population parameters. FAPS considers the probability of paternity of each offspring as a probability distribution over all candidate fathers simultaneously, and identifies and compares plausible sibling relationships. In this way we can fully account for the uncertainty in paternal and sibling relationships in a few seconds, obviating the need to update pedigree relationships iteratively. In this paper we describe how to include non-genetic information into the FAPS procedure, and use MCMC to update parameters for those data to jointly infer paternity, sibships and population parameters. We then apply this to the inference of the pollen dispersal distribution in a hybrid-zone population of the snapdragon *Antirrhinum majus*.

# 3 Materials and Methods

## 3.1 *A. majus* data

### 3.1.1 Study population

We examine a hybrid zone population of the snapdragon *Antirrhinum majus* in the Spanish Pyrenees. Here, the yellow-flowered *A. m. striatum* and the magenta-flowered *A. m. pseudomajus* meet and hybridise to produce diverse recombinant phenotypes, including pink, white and orange flowers. The population grows along two parallel roads running East-West close to Ribès de Freser (figure 1). The 'lower' road is at 1150-1200m above sea level, whilst the 'upper' road climbs 1250-1500m, and is 500-1000m north of the lower road. Hybrids are mostly confined to a 1km 'core' hybrid zone, with *A. m. striatum*- and *A. m. pseudomajus*-like plants becoming dominant to the West and East respectively. We surveyed as many flowering plants as we could find in June and July of 2012 (n=2124), and collected information on flower number and location using a Trimble GeoXT datalogger. We collected two to three leaves for DNA extraction and dried these in silica gel (Fischer Scientific). *A. majus* grows in disturbed habitats such as roadsides and railways; they are rare in the established forest and pasture between the two roads, on the north-facing slope to the South, and on the high mountain peak to the North of the two roads (figure 1. It thus is likely that we sampled the majority of the plants that flowered during the study period, although we cannot exclude that some flowering may have occurred before or after this.

Pollination is carried out exclusively by large bumblebees and carpenter bees who are large enough to open the flowers (Vargas et al., 2010; Andalo et al., 2019). *A. majus* has a gametophytic self-incompatibility system, and self-pollinated seeds are very rare (Surendranadh et al., 2022). There are no detectable post-zygotic barriers between *A. m. striatum* and *A. m. pseudomajus* (Andalo et al., 2010).

In August 2012 we collected a single, mature, wild-pollinated fruit from each of 60 mothers (figure 1). In order to minimise disturbance to the population we

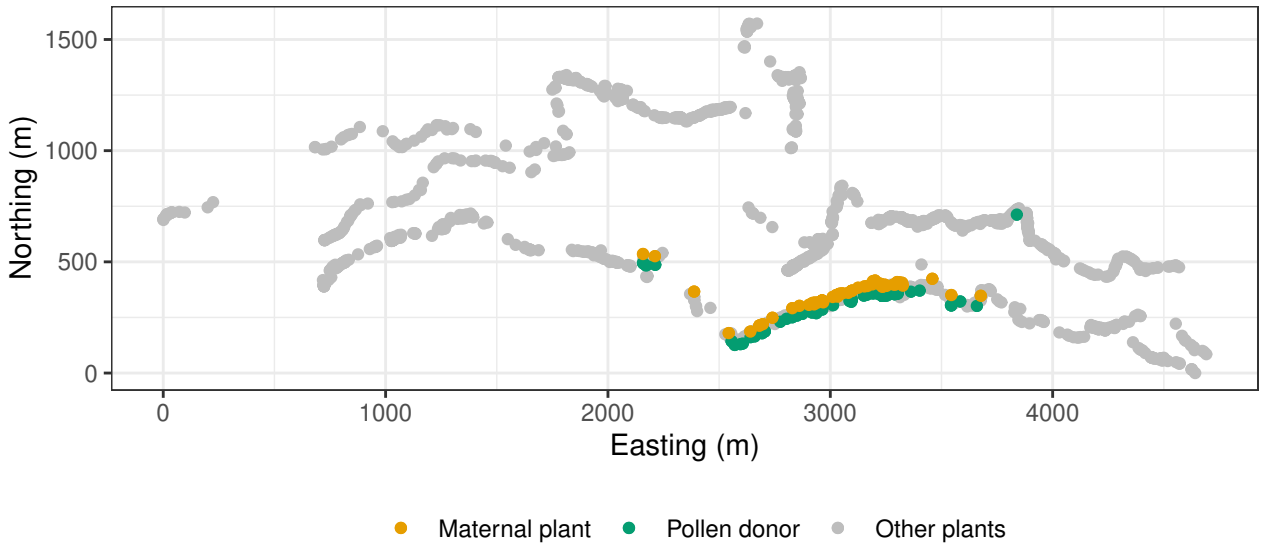


**Figure 1:** *Map of the hybrid zone. The map shows the distribution of maternal plants, inferred pollen donors and remaining plants along the lower (South) and upper (North) roads. Only pollen donors for families with one or more offspring are shown.*

only sampled from plants which had set a minimum of five mature fruits. These mothers were chosen to represent an even sample of pigmentation genotypes, spread as evenly as possible across the core of the hybrid zone where hybrids are most dense, resulting in 10 *A. m. striatum*-like, 17 *A. m. pseudomajus*-like, and 33 hybrid-phenotype mothers.

### 3.1.2 Genotyping

We grew seeds in 5cm plug trays filled with potting compost (Gramaflor) in a greenhouse under Sylvania GroLux lights on a 16-hour cycle. We sowed three seeds per plug for 50-70 plugs per maternal family and thinned seedlings to a single seedling per plug after cotelydons had appeared. We transferred approximately 1cm$^2$ of fresh tissue from 1419 seedlings to 96-well DNA-extraction plates (LGC Genomics, Berlin) and allowed tissue to dry using the sample bag and silica gel provided. For parental tissue from the hybrid zone we transferred approximately 1cm$^2$ tissue dried in the field to the same plates. DNA extractions of the plated tissue samples were carried out by LGC Genomics.

We genotyped tissue samples at 71 SNPs by KASPR sequencing (LGC Genomics). These SNPs are a subsample of a panel used for a wider survey of the hybrid zone (Surendranadh et al., 2022). The total SNP panel is a mixture of diagnostic (showing a gradient in allele frequency across the hyrbid zone) and parentage (with as even a gradient in allele frequency as possible) SNPs. Previous work identified the per-locus genotyping error rate in these data to be approximately $10^4$ (Surendranadh et al., 2022). For parentage loci we chose only biallelic loci with a minor-allele frequency greater than 0.3 in each of inner four pools closest to the centre of the cline, selected to be at least 2cM apart. Diagnostic SNPs were either linked to pigmenta-

tion loci, or else showed sharp clines across the hybrid zone. We removed 474 offspring and four adults that had missing data at more than 7.5% of the SNPs. We also pruned 7 SNPs that showed more than 10% missing data, or less than 15% heterozygosity. This left us with a set of 984 offspring from 60 maternal families, with between two and 29 offspring per maternal family (mean=16.4).

## 3.2 Joint estimation of paternity, sibships and dispersal

### 3.2.1 Probability model

We begin with observed SNP marker data **M** for mothers, offspring and candidate father, and a matrix **D** of Euclidean distances between mothers and all possible candidate fathers, as well as estimates $q$ of the proportion of candidate father sampled, and the per-locus genotyping error rate $\eta$. From this we wish to infer pedigree P describing sibling, paternal and (known) maternal relationships, and vector $\theta$ of dispersal parameters. The full probability model is then

$$\Pr(P, \theta | \mathbf{M}, \mathbf{D}, q, \eta) \propto \Pr(\mathbf{M} | P, \mathbf{D}, q, \theta, \eta) \Pr(\mathbf{D} | \theta, P) \Pr(\theta) \Pr(P) \Pr(\eta) \quad (1)$$

In the following sections we outline how to extend our existing method for inferring sibships and paternity to include data from non-genetic covariates, a model for pollen dispersal, suitable hyperprior distributions for dispersal parameters, and a procedure to infer the posterior distribution of the parameters of interest.

### 3.2.2 Allowing for covariates in paternity inference

We have previously described a Python package FAPS which performs joint analysis of paternity and sibship relationships for sibling arrays based on SNP data, and allows for integrating out uncertainty in relationships (Ellis, Field, and Barton, 2018). Here we extended the software to allow for additional non-genetic information to be included.

FAPS begins with marker data for one or more maternal families composed of a mixture of half- and full-siblings, the known mother of each family, and an array of candidate fathers. Ellis, Field, and Barton (2018) described a method to infer sibling and paternity relationships in those individuals; we refer the reader to that paper for the full details, but review the most important aspects here. FAPS uses marker data **M** to build matrix **G** of probabilities of paternity based on Mendelian transition probabilities for each maternal family. **G** has a row for each offspring and a column for each candidate father, where element $g_{ij}$ is the probability that candidate $j$ is the father of offspring $i$ given marker data, an estimate of genotyping error rate $\eta$, and an estimate of the proportion of missing candidate fathers $q$. Rows in **G** sum to one, and describe a multinomial distribution of probabilities of paternity over all candidate fathers. The final column of **G** is the probability that the father of each offspring was missing from the sample of candidates, based on the probability of observing offspring alleles from population allele frequencies, and an estimate of the proportion of possible pollen donors in the population that had been sampled. FAPS then builds a

similarity matrix whose $ih^{\text{th}}$ element is the likelihood

$$\Sigma_j^F g_{ij} g_{hj} \tag{2}$$

that the $i^{\text{th}}$ and $j^{\text{th}}$ offspring are full siblings by summing over the probabilities that they share any one of the $F$ fathers. The similarity matrix is used to perform hierarchical clustering to identify plausible ways to partition offspring into families of possible full sibships, and the likelihood of each partition structure is estimated by Monte-Carlo simulation. The likelihood that candidate father *j* is the father of putative full sibship *k* is then

$$\Pi_i^n g_{ij} \tag{3}$$

for all *n* offspring in *k*. FAPS returns a vector of probabilities (which sum to one) that each candidate is the father of sibship *k*, or that the true father was not sampled. This is done for each plausible partition structure. This gives a distribution of possible partition structures and their likelihoods, and a distribution of paternity probabilities for each putative family within each partition structure.

We can incorporate non-genetic information about paternity using a suitable function relating those data into probabilities of paternity. In principle this can be done for any kind of data for which there is a suitable function relating the observed data for a set of candidate fathers to a probability of having mating with each mother, given a set of parameter values for that function. The goal is to find parameter values that best explain the data. For example, a standardised continuous phenotype $z$ could be modelled with the logistic function $1/(1 + e^{\beta z})$, where $\beta$ describes the relationship between $z$ and male fertility. A categorical phenotype can be modelled as a multinomial vector of probabilities that sum to one. In this study we use matrix **D** of Euclidean distances between mothers and candidate fathers as a covariate, and model the probability $\Pr(d_{mj}|\theta)$ of a mating event occurring between mother *m* and candidate *j* who are distance $d_{mj}$ apart, based on a suitable model of pollen dispersal (see below). It is then straightforward to incorporate this into the procedure outlined above by modifying eqn. 2 to

$$\Sigma_j^F \Pr(d_{mj}|\theta) g_{ij} g_{hj} \tag{4}$$

and eqn. 3 to

$$\Pr(d_{mj}|\theta)\Pi_i^n g_{ij} \tag{5}$$

Eqn. 5 can then be used to estimate the likelihood of the partition structure by Monte-Carlo simulation, as previously described by Ellis, Field, and Barton (2018). For a given set of dispersal parameters in $\theta$ the likelihood of the whole maternal family is then the sum of likelihoods for each possible partition. When there are multiple maternal families, the likelihood of the whole dataset given $\theta$ is the product of those likelihoods over each maternal family. When we update $\theta$, this likelihood will change, giving us a way to compare likelihoods of different dispersal parameter values and identify those most consistent with the data.

A special case arises when it is known that offspring are unrelated, or that sibship structure is otherwise known. In this case the likelihood of the data can be calculated by summing the likelihoods of paternity for each candidate father on the $k^{th}$ full-sibship, and multiplying over all $K$ full sibships:

$$\Pi_k^K \Sigma_j^F \Pr(d_{mj}|\theta)\Pi_i^n g_{ij} \tag{6}$$

This obviates the need for likelihood estimation by Monte-Carlo simulation.

### 3.2.3 Pollen dispersal distribution

A useful function for describing plant dispersal distributions is the generalised normal distribution (Clark, 1998; Nadarajah, 2005; Kremer et al., 2012). This is a generalisation of the exponential family of probability distributions and includes the exponential and standard normal distributions as special cases, but allows for fat and thin tails. It is commonly used to model plant dispersal distributions because these are often found to show clear kurtotis (e.g. Austerlitz et al., 2004; Robledo-Arnuncio and Gil, 2005; Klein, Desassis, and Oddou-Muratorio, 2008; Burczyk, Sandurska, and Lewandowski, 2019; Field et al., 2011; Ottewell et al., 2012). The generalised normal distribution describes the probability of observing dispersal distance *d* given scale parameter $a$ and shape parameter $b$:

$$\Pr(d|a,b) = Ke^{(-\frac{d}{a})^b} \tag{7}$$

where K is normalising constant $b/[2a\Gamma(1/b)]$ and $\Gamma$ is Euler's gamma function. The function takes the forms of the standard exponential when $b = 1$ and normal distribution when $b = 2$. Values of $b < 1$ reflect leptokurtic distributions with tails that decay more slowly than would be expected under the exponential distribution. Using $\theta = a, b$, this provides a convenient function for calculating $\Pr(d_{mj}|\theta)$ because it allows for a long tail of long-distance migrants.

Sometimes it may be that an unrelated candidate has a similar or higher probability of paternity of one or more offspring simply due to stochasticity in Mendelian sampling. Whereas true fathers are expected to be (on average) close to the mother, other candidates should be drawn at random from the population. This will inflate the apparent kurtosis in the data and bias $b$ downwards. To accommodate this we modify 7 to model dispersal as a mixture of a generalised-normal and a uniform distribution:

$$\Pr(d_{mj}|a,b,\lambda) = \lambda Ke^{(-\frac{d}{a})^b} + \frac{1-\lambda}{F} \tag{8}$$

where $F$ is the number of candidate fathers and $\lambda$ is a mixture parameter determining the proportion of the probability mass due to 'real' dispersal. The uniform part of this mixture allows for signal coming from incorrect candidates without requiring the 'true' dispersal kernel to be unnecessarily leptokurtotic. $\lambda$ also provides an approximate estimate for what the rate of false-postive assignment to medium- and long-distance candidates would have been if this had been ignored.

It is common to also report the mean dispersal distance as the square root of the variance of this distribution as

$$\sqrt{\frac{a^2\Gamma(3/b)}{\Gamma(1/b)}} \tag{9}$$

(Nadarajah, 2005). We prefer to focus on median dispersal distances because this is more intuitive for highly skewed distributions, and also to estimate this on realised inter-mate distances rather than on parameters alone (see "Inference of mating patterns"). We nevertheless report this and how it relates to the effect of $\lambda$ on results.

### 3.2.4 Priors for dispersal parameters

We require prior distributions for dispersal parameters $a$, $b$ and $\lambda$.

We used a log-normal prior for $b$ ($\ln b \sim N(\mu = 0, \sigma = 0.5)$. This model allows for a range of shape values reflecting leptokurtosis to Gaussian dispersal, but is skeptical about very strong over- or under-dispersion

in the dispersal distribution because prior probabilties approach zero as $b$ approaches 0 or 3 (figure S1). We used a Gamma prior for $a$ because this describes positive continuous values, and because it is conjugate with the variance parameter of a Gaussian distribution. Because the effect of $a$ depends on $b$ and has no intuitive biological interpretation itself, we used prior simulations to choose a suitable parameterisation for $a$. We first simulated 10000 pairs of values for $a$ and $b$ from Gamma and log-normal distributions respectively. We then used each pair of simulated values to parameterise a generalised normal distribution, and simulated 1000 dispersal distances from that distribution, and visually examined the distribution for biological plausibility. We chose a Gamma distribution with shape=6 and scale=50, because this gave dispersal distributions with most dispersal occurring within 500m, but allowing for rare long-range dispersal events (figure S1).

For $\lambda$, we use a beta distribution with parameters Beta(1.1, 1.1). This distribution approaches zero when $\lambda$ is close to zero or one, but is fairly flat in between. This implies that we do not expect that all the weight should be on either the generalised-normal or the uniform components of the mixture distribution in eqn. 8, but that we do not have strong prior beliefs about values between that. To examine the effect of modelling dispersal as a mixture model at all, we also repeated the MCMC with with $\lambda$ to set to one.

### 3.2.5 Inference via MCMC

We used the Metropolis-Hastings Markov-chain Monte Carlo algorithm to infer the posterior distribution of dispersal parameters $a$,$b$ and $\lambda$. Ideally we would also like to estimate the posterior distribution of the proportion of missing fathers $q$, but we found this to be unstable. Moreover, we demonstrate by simulation that varying this parameter has negligible effect on biological conclusion (see below). We therefore used a fixed value of $q = 0.5$, which is tantamount to a flat prior on sampling effort.

We ran four independent chains beginning from distant areas of the parameter space. At each iteration, we perturbed the value of each parameter by a factor drawn from a normal distribution with a fixed standard deviation for each parameter ($\sigma = 2$ for $a$; $\sigma = 0.05$ for $b$; $\sigma = 0.025$ for $\lambda$). We ran each chain for 40000 iterations and subsequently removed the first 500 iterations of each as burn-in. After checking chain convergence we thinned subsequent iterations to retain 250 posterior draws from each chain for further analyses, for a total of 1000 posterior draws.

### 3.2.6 Inference of mating events

We aimed to create a list of possible mating events between mothers and candidate fathers consistent with the data to identify a set of independent pollen dispersal events. For each partition structure, FAPS identifies valid set of distinct fathers for each full sibship, and estimates a likelihood for each father-sibship configuration. Each set of fathers reflects a hypothetical set of mating events, and the probability that a single father mated with the maternal plant is the sum of probabilities for each partition structure in which he sires at least one offspring. This gives a list of mating events, with a posterior probability that each occurred, including an entry for offspring whose father was missing. The posterior estimate of offspring number is the number of offspring for each father in each father-sibship configuration weighted by the probability of the mating event. Note that the number and size of sibships are not neces-

sarily integers because estimates are weighted averages over plausible sibship partition structures. In particular, hypothetical families can have posterior sizes less than one if the nominal number of offspring is one, but the posterior probability for the family is less than one.

The total number of mating events is the sum of probabilities for all mating events with sampled fathers. Likewise, an estimate of the number of missing fathers is the sum of probabilities for all mating events for unsampled fathers. However, FAPS does not distinguish paternal families within groups of offspring of an unsampled father, so this value is a lower bound on the number of unsampled fathers.

For 1000 iterations of the MCMC output we inferred mating events in this way, based on genetic information and probabilities of dispersal from the scale and shape values for that iteration in sibship clustering (eqn. 4 and 3). This gives 1000 sets of mating events for each iteration of the MCMC output.

Finally, we estimated the relationship between the number of offspring in a maternal array and the number of paternal families therein. We first averaged paternal family size for each mother-father pair over iterations of the MCMC, then fitted a Poisson generalised linear model (GLM) of paternal family size on maternal family size (McCullagh and Nelder, 1989).

## 3.3 Power analysis

We used simulations to determine the statistical power of the dataset and method to identify true fathers. We simulated mating events based on pollen dispersal from 200 draws from the posterior distributions of shape and scale parameters. For each iteration in the chain we simulated mating events between each true mother and one or more sires drawn from the pool of genotyped adult plants. We drew sires based on the probability of mating with a mother given dispersal parameters for that iteration. For each mother, we simulated as many mating events of the same size as were inferred in that mother in the observed data. We ignored mating events with posterior probabilities less than 0.9. We simulated offspring genotypes based on Mendelian segregation and added genotyping errors to adults and offspring genotypes with the observed genotype error rate of $10^{-4}$ per locus per individual. In this way simulated data reflects the structure of the empirical dataset.

We also investigated the relationship between the true proportion of missing fathers and prior assumptions about sampling effort. To do this we randomly removed 10%, 30% or 50% of the true fathers for each simulated dataset. We then used FAPS to infer mating events for each dataset using prior probabilties of the proportion of missing fathers of 10%, 20%, 30%, 40% and 50% and counted how often we were able to recover correct paternal families, the number of incorrect paternal families and the number of offspring assigned to missing fathers. This simulates the effect of incomplete sampling, as well as of over- and under-estimating sampling effort subsequent analysis.

## 3.4 Asymmetric dispersal

To test for potential bias in the direction of pollen dispersal we calculated the mean difference in East-West position between mothers and fathers. We compared the average asymmetry in direction for observed mating events to simulated mating events as described above. Because all but one observed pollen donor turned out to be on the lower road we restricted this analysis to simulated pollen donors on the lower road.

# 4 Results

## 4.1 True sibships can be identified with high confidence

We found that paternal families inferred from simulated data could be divided into those with family sizes of one or more, and a second group with weighted-mean family sizes less than one (figure 2A). Consistent with this, families with one or more offspring overwhelmingly reflect true mating events, whereas those with less than one offspring were overwhelmingly incorrect (figure 2B). This indicates that we can reliably identify true mating events by discarding paternal families with weighted-mean family sizes less than one.

We found surprisingly little effect of prior beliefs about the proportion of missing fathers on the accuracy of paternity inference. Focussing on robust paternal families with one or more offspring, the number of families detected decreased as the true number of missing fathers increased, as one would expect (figure 2A). However, there was negligible relationship between the number of these families found and the prior expectation about the proportion of missing fathers, even when this prior substantially over- or under-estimated the true value. We did find a slight increase in the probability that these inferred mating events were correct with higher input values of $q$ (figure 2B), presumably because increasing the prior probability that a father is unsampled decreases the probability of paternity for candidates who are not true fathers. Likewise, the number of offspring assigned to missing fathers increased with the true proportion of missing fathers, but there was no effect of changing the value of proportion of missing fathers used in the analysis (figure 2C). These results demonstrate that the prior expectations about the proportion of missing fathers has very little effect on downstream analyses.

We also note that the observed number of offspring with an unsampled father in the simulations is substantially higher than would be expected from the true proportion of fathers. This is because the simulations use the true spatial distribution of maternal and candidate plants, and the same candidate can sire offspring with multiple mothers. If that candidate is unsampled, this affects multiple full-sib families. As such, assuming our simulation regime is sufficiently realistic, we should expect the proportion of offspring with an unsampled father in the empirical dataset to be higher than the true proportion of missing fathers.

## 4.2 Number and size of full-sibling families

We identified an average of 160.6 full-sibling families for which a father could be positively identified across MCMC samples (96% credible intervals (CI): 156, 163). Consistent with simulation results, 18.6 families had posterior-mean family sizes less than one (96% CIs: 14, 121; figures S2, S3). These families typically had posterior probabilities less than one and showed substantial uncertainty across MCMC iterations, indicating that they were detected only for a subset of possible sibship configurations (figure S2). In contrast, the remaining 142.0 families with at least one offspring had posterior probabilities greater than 0.99 and were found across MCMC iterations (figure S2). These paternal families likely reflect robust independent mating events that can be used to infer a dispersal kernel, and we focus on these mating events in the rest of this manuscript.

Full-sibship sizes ranged from one to 15 (figure S3). The GLM of the number of full sibships on maternal-

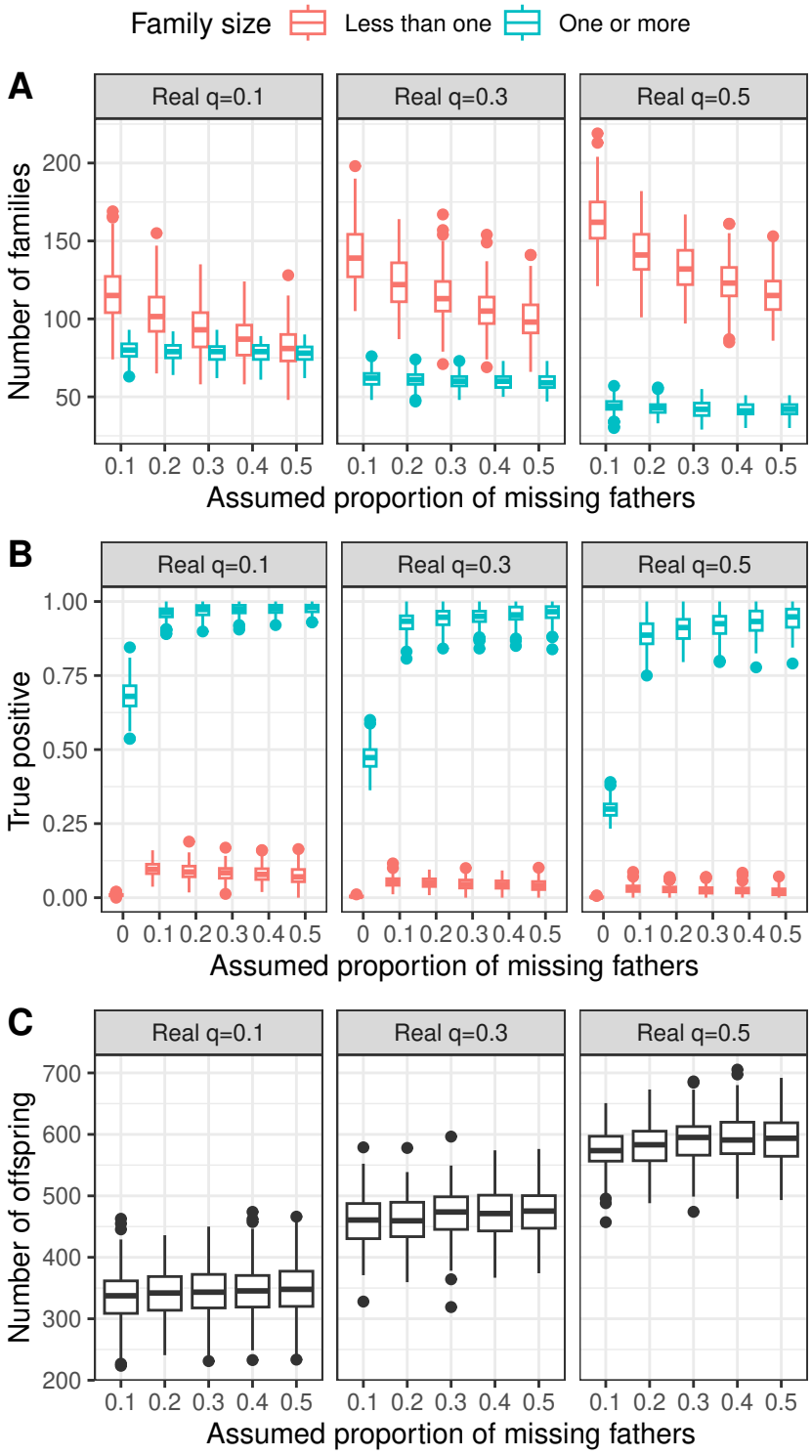


**Figure 2:** *Simulation results. Panels show the true proportion, q, of missing fathers; x-axes show the assumed proportion used to infer families. (A) Number of paternal families identified with weighted-mean offspring number averaged over plausible sibships greater or less than one. (B) Probability that paternal families identified by FAPS are real. (C) Number of offspring inferred to have an unsampled father.*

family size revealed that (natural) log number of sibships increased by 0.066 ($\pm$0.015) for every additional offspring included in the maternal family (intercept = -0.173 $\pm$ 0.307; figure S4). This corresponds to detecting a new family for approximately every 3 or 4 offspring genotyped, on average.

For all sixty mothers we found a mating event with non-zero probability for which the father one or more offspring was not sampled (figure S3). These families include an average of 590.2 offspring (CIs: 588.3, 592.3). This corresponds to around 60% of the offspring, and a minimum of 28.1% (CIs 27.8, 28.3) of the total sample of pollen donors.

## 4.3 Pollen dispersal is leptokurtic and asymmetric

Across prior dispersal scenarios, independent MCMC chains converged on the stationary distribution and were generally well mixed (figures 3, S5, S6). The posterior distribution for the disperal shape parameter was consistently less than one, with a posterior mean of 0.56 (96% CIs: 0.46, 0.70), indicating a leptokurtic dispersal kernel (figure 3B). Consistent with this, the distribution of pollen dispersal distances for families with one or more offspring shows a peak close to zero and a long tail of long-distance dispersal events up to 747m (figures 4A, S7). Mean and median dispersal distances were 78.0 and 26.7 metres respectively. With one exception, all mating occured between plants on the lower road (figure 1). There was no evidence for a relationship between the posterior probability of a mating event and the distance to the father for these mating events (figure S7). These results indicate a leptokurtic pollen dispersal kernel.

Pollen donors were on 26% more likely to be to the East of the maternal plants they mated with than to the West. Simulated pollen donors were 5.1% more likely to be to the East of their mates, indicating that the observed result is at least partially due to increased density of plants to the East of the sample of maternal plants. Nevertheless the average position of observed pollen donors was 36.7 metres further to the East than their mates, which is further than in 98.0% of simulated datasets (mean distance = 12.8 metres; 96% CIs: -5.93, 32.7). This indicates that there is a bias in pollen flow favouring Westward dispersal.

## 4.4 Mixture parameter reduces bias in dispersal shape

The posterior distribtion the mixture parameter $\lambda$ was centred around 0.93 (96% CIs: 0.88, 0.96; figure 3C). That this number is close to, but not equal to zero indicates that 4-12% of paternal families are compatible with non-sires located at medium to long distances from the mother. This may be because the true father is missing and/or stochastictity in Mendelian sampling of the genotypes. If ignored, this signal would inflate apparent leptokurtosis in the dataset. Consistent with this, when we constrain $\lambda$ to be fixed at one the posterior distributions for the scale and shape parameters are shifted downwards compared to the analysis where $\lambda$ is allowed to vary (posterior means for $a$ and $b$ drop from 17.5 to 17.1 and from 5.6 to 4.9 respectively; figure S8). This would indicate more leptokurtic pollen dispersal (figure 3A, 3B). When mean dispersal distance is estimated from the second moment of the generalised normal distribution (eqn. 9; Clark, 1998) this would imply a substantial increase in mean dispersal from 114 to 193 metres (figure S8). Fortunately, there was no difference in mean or median dispersal distance when

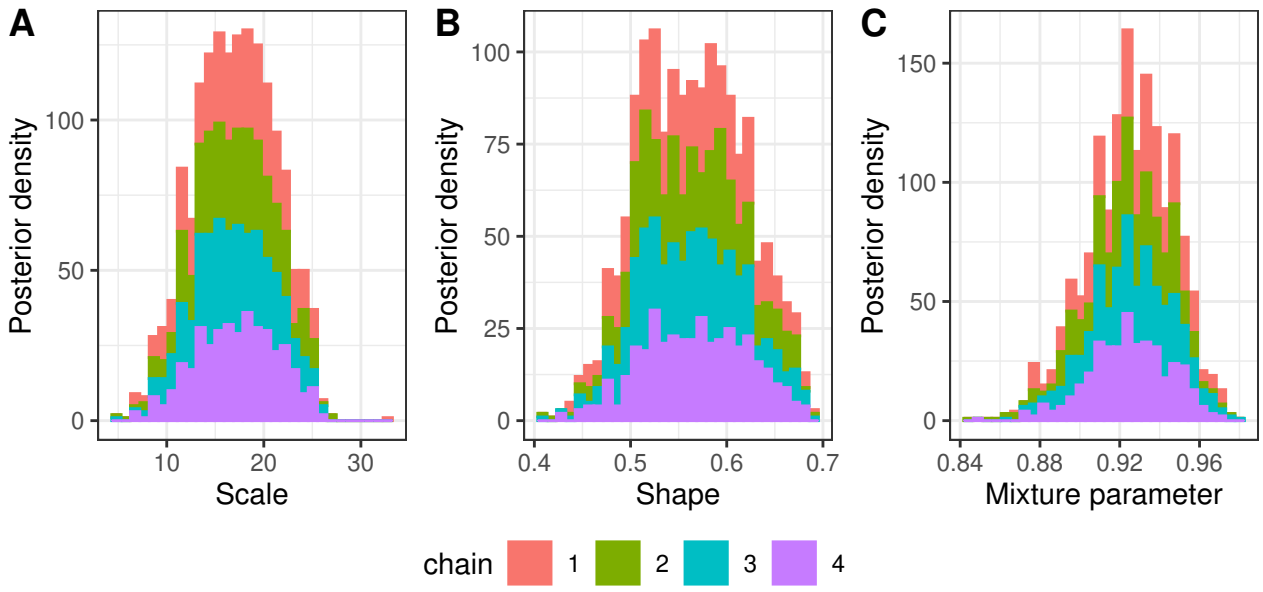


**Figure 3:** *Posterior densities for the scale, shape, and mixture parameters of the dispersal kernel. Histograms show stacked densities for each of four independent chains.*

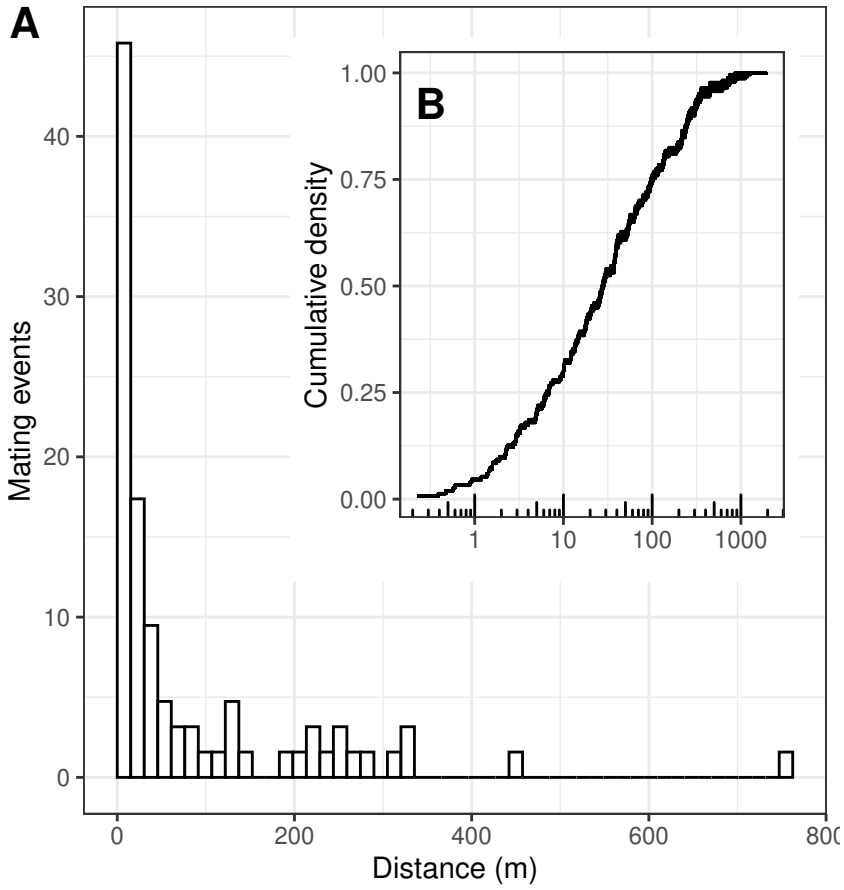


**Figure 4:** *Distribution of pollen-dispersal distances. (A) Histogram and (B) cumulative distribution of distances for dispersal events for paternal families with one or more offspring. Note the log10 scale of the x-axis in B. Mating events are weight weighted by their posterior probability. The histogram is summed over iterations of the MCMC. Separate cumulative curves are shown for 1000 MCMC iterations separately.*

these are calculated based on realised mating events. In this instance then, including a mixture component in the dispersal kernel reduces the bias caused by missing fathers in dispersal shape and apparent mean dispersal distance when this is inferred from the shape of the dispersal distribution. However, this does not have a large enough impact on realised dispersal distances in these data to greatly affect biological conclusions.

# 5 Discussion

We have described a framework to jointly infer paternity, sibling relationships and population parameters. Applying this to a natural populations of snapdragons, we find strong evidence for a leptokurtic pollen-dispersal distribution, as well as evidence for asymmetry in pollen dispersal from East to West. Below we discuss possible influence of incorrect genealogies on the results, the implications for the hybrid zone population, as well as the limitations and possible future directions of the method.

## 5.1 Controlling bias in dispersal estimates

### 5.1.1 Bias due to incorrect fathers

Sometimes candidate pollen donors can have a similar or greater probability of paternity for an offspring individual as does the true father due to stochasticity in Mendelian sampling (Thompson, 1976). If not addressed, we might infer an incorrect mating event between adult plants, which would bias estimates of dispersal upwards. Two aspects of our study suggest that signal from false fathers should have a minimal effect on our results.

First, our simulations showed that incorrect families could be readily identified based on family size. In previous simulations to test the performance of FAPS we found that the main source of error was individuals from a larger sibship group being falsely assigned to a singleton sibship, but this analysis focussed on the most likely partition structure to simplify results (Ellis, Field, and Barton, 2018). In this study we found that those incorrect singleton families have only weak posterior probabilities when we consider a whole set of plausible sibship configurations, meaning that family sizes are effectively less than one individual. This idea is not necessarily intuitive, so to illustrate this, imagine three offspring (A, B and C) with two candidate fathers (X and Y) and that there are two plausible paternal-sibships configurations,: (1) X is the father of A, B and C, with probability 0.6, or (2) X is the father of A and B, while Y is the father of C with probability 0.4. The probability that X mated with the mother is 0.6+0.4=1, but the probability that Y mated with the mother is 0.4. Weighted mean offspring numbers are (3*0.6)+(2*0.4)=2.4 for X and 1*0.5=0.5 for Y. Since our goal was to infer pollen dispersal by identifying mate pairs, excluding mating events for families with sizes less than one is an effective way to control false-paternity assignment. This demonstrates that accounting for uncertainty in sibship structure is an extremely useful way to improve the accuracy of paternity inference.

Second, we modelled dispersal as a mixture distribution with a term allowing for false-positive fathers to try and reduce the bias in estimates of dispersal. The mixture distribution reduces the bias in the estimated shape parameter and the second moment of the generalised normal distribution, often interpreted as mean (squared) dispersal distance (e.g. Clark, 1998; Auster-

litz et al., 2004; Klein, Desassis, and Oddou-Muratorio, 2008). We note that this bias will exist to some extent as long as knowledge of paternity is less than perfect. As such the mixture model reduces this bias, but is unlikely to completely eliminate it completely. Our biological conclusions are not affected, because we focus on the distribution of realised mating events, which are affected in only a minor way in real terms (figure 4C, 4D). However, caution is warranted in the interpretation of the raw parameter values for shape and mean dispersal from the generalised normal distribution (figure S8). It is possible that the effect was stronger here than would be the case in other published studies because our sample of candidate fathers is large, and the spatial extend is much larger (two orders of magnitude) than median dispersal distance. In contrast, other studies have focussed on wind-pollinated trees, where pollen can travel much further, and had smaller numbers of candidates (e.g. Adams, Griffin, and Moran, 1992; Austerlitz et al., 2004; Klein, Desassis, and Oddou-Muratorio, 2008). This demonstrates the utility of using inferred dispersal parameters to inform inference of real mating events, rather than focussing on the parameters themselves.

### 5.1.2 Missing fathers

An alternative source of bias comes from false negative paternity, where the true father of a family is present but is inferred to be missing. We found that more than half the offspring were inferred to have an unsampled father, and that this did not strongly depend on prior beliefs about the proportion of missing fathers. Whether these missing fathers bias dispersal estimates depends on whether they are likely to be randomly distributed in space. On one hand, it may be that sampling effort was greater in the core of the hybrid zone than at the edges of our sampling region, and it is known that bumblebees, especially queens, can disperse over several kilometres (Osborne et al., 2008; Hagen, Wikelski, and Kissling, 2011; Lepais et al., 2010). In this way it is possible that we have missed long-distance mating events, which would bias dispersal estimates downward. However, our sample of candidate fathers is large, and extends at least a kilometer beyond the range of plants identified as pollen donors (figure 1), which is more than twice as far as the longest dispersal event detected (figure 4). We feel it is thus unlikely that there could have been enough unsampled fathers far from the sample of mothers to greatly influence dispersal shape or median dispersal distance.

An alternative explanation for the high number of missing fathers is that they finished flowering before the main sampling began. This is a more plausible explanation, because the population had already begun to flower at the time of the first surveys in June 2012, and plants were still flowering at the time of seed collection in late July. The majority of unsampled fathers may thus represent relatively early-flowering individuals. Fortunately, it seems that flowering time is unlikely to depend strongly on location, and as such this is unlikely to bias dispersal estimates. Ongoing work using many more samples of mothers across both space and time will allow us to address questions such as this.

## 5.2 Implications for the hybrid-zone population

The shape of the dispersal kernel implies that most mating occurs within tens of metres, but that there is a long tail of mating events between individuals at up

to several hundred metres. This result suggests that any natural selection acting through male reproductive components (pollen dispersal) occurs at small spatial scales. If pollinators discriminate between different flower-colour phenotypes in the hybrid zone, this sets an important geographic scale of selection for empirical tests on pollinator-mediated selection. For example, within 20m in the core of the hybrid zone, pollinators would encounter a full range of colour phenotypes that would provide scope for colour discrimination and differential visitation to influence male and female fitness. In this paper we have not attempted to estimate selection via flower colour because the number of mothers for each flower colour is small, meaning that we likely have little statistical power. However, in ongoing work we are investigating variation in fitness in a much larger panel of offspring arrays, which will allow us to address such questions with much greater accuracy.

We found evidence that all but one mating event was between plants on the lower road, and that pollen donors on this road tended to be to the East of maternal plants. This indicates that pollen tends to move Westward from where the magenta *A. m. pseudomajus* dominates towards where yellow *A. m. striatum* is more common, implying an asymmetry of introgression from *A. m. pseudomajus* into *A. m. striatum*. We note however that our sample of maternal plants was restricted to a relatively narrow region around the core of the hybrid zone, to only one of the two roads where *A. majus* is common (figure 1), and in only a single year. As such, the finding of asymmetric introgression should be regarded as preliminary. In ongoing work we are investigating the detailed shape of clines across the genome in the population that reflect demographic processes over longer time scales, and should be able to shed clearer light on this phenomenon.

## 5.3 Limitations and future directions

In this study we used Metropolis-Hastings MCMC to update population parameters, and re-evaluate sibship structures at each iteration using the Monte-Carlo simulations described by Ellis, Field, and Barton (2018). This has two obvious computational limitations that could be improved upon. First, the Metropolis-Hastings algorithm known to explore parameter space inefficiently as models become complex. However, more efficient MCMC samplers, such as Hamiltonian Monte Carlo are challenging to implement, especially when the likelihood component needs to be estimated by simulation (Betancourt, 2017). Second, the Monte-Carlo simulations take several seconds to run, which means chains with many iterations take a long time.

These issues could be greatly simplified by avoiding repeated Monte-Carlo simulations at each iteration. One way to achieve this would be to first identify full-sibling families with strong support based on genetic data Ellis, Field, and Barton, 2018). Our simulations showed that this can be done with high confidence based on family size. Then, these families could be used to jointly infer paternity and population parameters following eqn. 6. In this case the likelihood is the sum of elements in a matrix, which is very cheap to compute. This would allow for much more flexibility in how to optimise parameter estimation. This is turn would allow users to focus less on implementation and more on modelling biological problems.

## 6 Acknowledgements

We thank a large number of field volunteers for maintaining the population sampling, and Tom White for assistance with seed collection. We thank Sylvia Rebel for plating tissue for DNA extraction and Sean Stankowski for feedback on the manuscript.

## 7 Data availability

Data and code to recreate the analyses are available from Zenodo (DOI: DOI: 10.5281/zenodo.10462674). A supplemental file listing inferred mating events is included as a supplementary file.

## Bibliography

Adams, WT, AR Griffin, and GF Moran (1992). "Using paternity analysis to measure effective pollen dispersal in plant populations". In: *The American Naturalist* 140.5, pp. 762–780.

Andalo, C et al. (2010). "Post-pollination barriers do not explain the persistence of two distinct *Antirrhinum* subspecies with parapatric distribution". In: *Plant Systematics and Evolution* 286.3, pp. 223–234.

Andalo, Christophe et al. (2019). "Prevalence of legitimate pollinators and nectar robbers and the consequences for fruit set in an Antirrhinum majus hybrid zone". In: *Botany Letters* 166.1, pp. 80–92.

Anderson, Eric C and Thomas C Ng (2016). "Bayesian pedigree inference with small numbers of single nucleotide polymorphisms via a factor-graph representation". In: *Theoretical Population Biology* 107, pp. 39–51.

Austerlitz, Frederic et al. (2004). "Using genetic markers to estimate the pollen dispersal curve". In: *Molecular ecology* 13.4, pp. 937–954.

Betancourt, Michael (2017). "A conceptual introduction to Hamiltonian Monte Carlo". In: *arXiv preprint arXiv:1701.02434*.

Bullock, James M et al. (2017). "A synthesis of empirical plant dispersal kernels". In: *Journal of Ecology* 105.1, pp. 6–19.

Burczyk, Jarosław, Elżbieta Sandurska, and Andrzej Lewandowski (2019). "Patterns of effective pollen dispersal in larch: linking levels of background pollination with pollen dispersal kernels". In: *Forests* 10.12, p. 1139.

Cain, Michael L, Brook G Milligan, and Allan E Strand (2000). "Long-distance seed dispersal in plant populations". In: *American journal of botany* 87.9, pp. 1217–1227.

Clark, James S (1998). "Why trees migrate so fast: confronting theory with dispersal biology and the paleorecord". In: *The American Naturalist* 152.2, pp. 204–224.

Ellis, Thomas James, David Luke Field, and Nicholas H Barton (2018). "Efficient inference of paternity and sibship inference given known maternity via hierarchical clustering". In: *Molecular Ecology Resources* 18, pp. 988–999. DOI: 10.1111/1755-0998.12782.

Emery, AM et al. (2001). "Assignment of paternity groups without access to parental genotypes: multiple mating and developmental plasticity in squid". In: *Molecular Ecology* 10.5, pp. 1265–1278.

Field, David L et al. (2011). "The importance of premating barriers and the local demographic context for contemporary mating patterns in hybrid zones

of *Eucalyptus aggregata* and *Eucalyptus rubida*". In: *Molecular Ecology* 20.11, pp. 2367–2379.

Hadfield, JD, DS Richardson, and T Burke (2006). "Towards unbiased parentage assignment: combining genetic, behavioural and spatial data in a Bayesian framework". In: *Molecular Ecology* 15.12, pp. 3715–3730.

Hagen, Melanie, Martin Wikelski, and W Daniel Kissling (2011). "Space use of bumblebees (*Bombus* spp.) revealed by radio-tracking". In: *PloS one* 6.5, e19997.

Huisman, Jisca (2017). "Pedigree reconstruction from SNP data: parentage assignment, sibship clustering and beyond". In: *Molecular ecology resources* 17.5, pp. 1009–1024.

Jones, Beatrix et al. (2007). "Estimating differential reproductive success from nests of related individuals, with application to a study of the mottled sculpin, Cottus bairdi". In: *Genetics* 176.4, pp. 2427–2439.

Klein, Etienne K, Nicolas Desassis, and Sylvie Oddou-Muratorio (2008). "Pollen flow in the wildservice tree, Sorbus torminalis (L.) Crantz. IV. Whole interindividual variance of male fecundity estimated jointly with the dispersal kernel". In: *Molecular Ecology* 17.14, pp. 3323–3336.

Kremer, Antoine et al. (2012). "Long-distance gene flow and adaptation of forest trees to rapid climate change". In: *Ecology letters* 15.4, pp. 378–392.

Lepais, Olivier et al. (2010). "Estimation of bumblebee queen dispersal distances using sibship reconstruction method". In: *Molecular Ecology* 19.4, pp. 819–831.

McCullagh, Peter and JA Nelder (1989). *Generalized Linear Models*. 2nd ed. New York: Chapman and Hall.

Nadarajah, Saralees (2005). "A generalized normal distribution". In: *Journal of Applied Statistics* 32.7, pp. 685–694.

Neff, Bryan D, Joe Repka, and Mart R Gross (2001). "A Bayesian framework for parentage analysis: the value of genetic and other biological data". In: *Theoretical Population Biology* 59.4, pp. 315–331.

Osborne, Juliet L et al. (2008). "Bumblebee flight distances in relation to the forage landscape". In: *Journal of animal ecology* 77.2, pp. 406–415.

Ottewell, K et al. (2012). "The pollen dispersal kernel and mating system of an insect-pollinated tropical palm, Oenocarpus bataua". In: *Heredity* 109.6, pp. 332–339.

Pemberton, JM (2008). "Wild pedigrees: the way forward". In: *Proceedings of the Royal Society B: Biological Sciences* 275.1635, pp. 613–621.

Robledo-Arnuncio, JJ and L Gil (2005). "Patterns of pollen dispersal in a small population of Pinus sylvestris L. revealed by total-exclusion paternity analysis". In: *Heredity* 94.1, pp. 13–22.

Surendranadh, Parvathy et al. (2022). "Effects of fine-scale population structure on the distribution of heterozygosity in a long-term study of *Antirrhinum majus*". In: *Genetics* 221.3, iyac083.

Thomas, Stuart C and William G Hill (2002). "Sibship reconstruction in hierarchical population structures using Markov chain Monte Carlo techniques". In: *Genetics Research* 79.3, pp. 227–234.

Thompson, EA (1976). "A paradox of genealogical inference". In: *Advances in Applied Probability* 8.4, pp. 648–650.

Vargas, Pablo et al. (2010). "Is the occluded corolla of Antirrhinum bee-specialized?" In: *Journal of Natural History* 44.23-24, pp. 1427–1443.

975 Wang, J (2007). "Parentage and sibship exclusions:
976 higher statistical power with more family members".
977 In: *Heredity* 99.2, pp. 205–217.
978 Wang, Jinliang (2004). "Sibship reconstruction from
979 genetic data with typing errors". In: *Genetics* 166.4,
980 pp. 1963–1979.