Received: 28 May 2019 | Revised: 27 July 2019 | Accepted: 19 August 2019

DOI: 10.1111/1755-0998.13094

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

Individualized mating system estimation using genomic data

Jack Colicchio | Patrick J. Monnahan | Carolyn A. Wessinger | Keely Brown | James Russell Kern | John K. Kelly

Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, KS, USA

Correspondence
John K. Kelly, Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, KS, USA.
Email: jkk@ku.edu

Present addresses
Jack Colicchio, Department of Plant and Microbial Biology, University of California Berkeley, Berkeley, CA, USA

Patrick J. Monnahan, Plant and Microbial Biology, University of Minnesota, St. Paul, MN, USA

Funding information
National Institutes of Health, Grant/Award Number: R01 GM073990

Abstract

The estimation of outcrossing rates in hermaphroditic species has been a major focus in the evolutionary study of reproductive strategies, and is also essential for plant breeding and conservation. Surprisingly, genomics has thus far minimally influenced outcrossing rate studies. In this article, we generalize a Bayesian inference method (BORICE) to accommodate genomic data from multiple subpopulations of a species. As an empirical demonstration, BORICE is applied to 115 maternal families of *Mimulus guttatus*. The analysis shows that low-level whole genome sequencing of parents and offspring is sufficient for individualized mating system estimation: 208 offspring (88.5%) were definitively called as outcrossed, 23 (9.8%) as selfed. After mating system parameters are established (each offspring as outcrossed or selfed and the inbreeding level of maternal plants), BORICE outputs posterior genotype probabilities for each SNP genomewide. Individual SNP calls are often burdened with considerable uncertainty and distilling information from closely linked sites (within genomic windows) can be a useful strategy. For the Mimulus data, principal components based on window statistics were sufficient to diagnose inversion polymorphisms and estimate their effects on spatial structure, phenotypic and fitness measures. More generally, mating system estimation with BORICE can set the stage for population and quantitative genomic analyses, particularly researchers collect phenotypic or fitness data from maternal individuals.

KEYWORDS

BORICE, Mimulus, outcrossing, RAD-seq, selfing

1 | INTRODUCTION

The mating system of a hermaphroditic species is determined by how frequently individuals self-fertilize and the pattern of matings when they outcross. These mating system parameters have pronounced ecological and evolutionary effects (Eckert et al., 2010; Pannell & Voillemot, 2017). Selfing can allow populations to persist through intervals of low pollinator abundance (Kalisz & Vogler, 2003; Müller, 1883) and also to establish in novel habitats (Baker, 1955; Grossenbacher, Briscoe Runquist, Goldberg, & Brandvain, 2015). It can also restrict gene flow across a metapopulation (Duminil, Hardy, & Petit, 2009; Govindaraju, 1988) and reduce progeny fitness through inbreeding depression. Within populations, the mixture of outcrossing and selfing determines the joint distribution of genotypic and phenotypic variation (Cockerham & Weir, 1984; Kelly, 1999a, 1999b; Kempthorne, 1957; Wright, 1951), and thus the response to selection, either natural or artificial.

Mating system experiments typically sample maternal plants or animals from one or more natural populations, along with their progeny. Progeny, and sometimes the maternal parent, are genotyped at a set of putatively neutral markers. Comparison of progeny genotypes to maternal genotypes (directly scored or inferred) is then used to estimate *t*, the mean population outcrossing rate, and *F*, the mean inbreeding coefficient of maternal individuals (Ritland, 1990, 2002; Ritland & Jain, 1981). These population level estimates (*t* and *F*) have been obtained from hundreds of species and brought to bear on questions in ecology, evolution, and conservation (Eckert et al., 2010; Goodwillie, Kalisz, & Eckert, 2005; Whitehead, Lanfear, Mitchell, & Karron, 2018).

An important challenge is to estimate mating system parameters at the scale of individuals – to call individual offspring as outcrossed or selfed and determine whether specific maternal individuals are inbred and to what extent. With progeny scored definitively, we can investigate how selfing rate varies among individuals in a population, and subsequently the genetic, phenotypic, and ecological determinants of this variation. Genetic details such as allelic effect size and dominance are important for the evolution of selfing (Holsinger, 1988; Lande & Schemske, 1985; Uyenoyama, Holsinger, & Waller, 1993), but we remain largely ignorant about these details, particularly for variants that segregate within populations. There is great interest in the evolvability of selfing as climate change is decoupling flowering from pollinator occurrence (Forrest, 2015; Gezon, Inouye, & Irwin, 2016; Kudo & Ida, 2013). In this article, we update Bayesian outcrossing rate inbreeding coefficient estimation (BORICE) (Koelling, Monnahan, & Kelly, 2012) to determine whether offspring are outcrossed (δ = 0) or selfed (δ = 1) using genomic data. As an empirical application, we use these individual estimates to test whether flowers produced early in the season differ in selfing rate from those produced later, as might be expected given changes in pollinator availability.

With adults (maternal individuals) scored definitively, we can address questions about inbreeding depression in nature. An important question for many mixed mating species is whether any selfed progeny survive to reproduce. Reviewing estimates for *F*, Scofield and Schultz (2006) found that while both short and long-lived plant species self-fertilize, only the former seem to produce reproductive adults through selfing (*F* > 0). Inbreeding depression is apparently sufficiently severe in long-lived species that selfed seed, even if produced frequently, never make it to maturity. This total failure hypothesis is surprising from an adaptationist viewpoint: why invest in producing seeds that will never succeed? *F* estimates indistinguishable from zero are certainly consistent with total failure, but strong conclusions about inbred plant fitness require inference of individual inbreeding histories. Infrequent survival of inbred plants, say 1 in 100 or 1 in 1,000, may be adaptively sufficient to maintain delayed selfing (Lloyd, 1979), but will have a minimal effect on *F* (which is a population mean value). BORICE estimates *F* in terms of the inbreeding history (*IH*) of each individual adult. We treat the adult population as a series of discrete cohorts defined by IH (Campbell, 1986; Kelly, 1999b). *IH* = 0 for outbred individuals, each of which has an inbreeding coefficient, *f*, equal to 0. Selfed progeny of outbred individuals (*f* = ½) have *IH* = 1, while second generation selfs (the selfed progeny of *IH* = 1 individuals) have *IH* = 2 (*f* = ¾), and so on. In principle, $IH \rightarrow \infty$, but we bin all *IH* ≥ 7 into one final category (*f* > 0.99). The total failure hypothesis is rejected by field data if the posterior probability that *IH*> 1 approaches one for even a few adults in the population.

With an appropriate analytical framework, genome-wide genotyping can provide individual level resolution of mating system parameters. Surprisingly however, while genomic methods have been applied for paternity inference (Ellis, Field, & Barton, 2018; Flanagan & Jones, 2019), they have not been widely used for outcrossing rate estimation. Investigators still primarily use PCR based markers such as microsatellites, e.g., Ladd, Thavornkanlapachai, & Byrne, 2018; Mignot et al., 2016; Nagamitsu, Shuri, Taki, Kikuchi, & Masaki, 2016; Nishizawa & Ohara, 2018; Palma-Silva, Cozzolino, Paggi, Lexer, & Wendt, 2015; Rivière et al., 2019; Robertson et al., 2015; Yang et al., 2018. Techniques like RAD-seq provide a 100-fold increase in genetic data per individual (at minimum) for the same effort/investment, but their use in estimating mating system parameters is hindered by difficulties in scaling up to analyze thousands of loci and also by higher per-locus error rates. BORICE accommodates uncertainty of single nucleotide polymorphism (SNP) specific calls by integrating evidence across the genome.

Figure 1 is an overview of the BORICE pipeline as applied to an experiment investigating yellow monkeyflower, *M. guttatus*. The first outputs are the mating system parameters (IH for each parent, δ for each offspring) and allele frequencies within each of multiple subpopulations. With IH and δ established (estimated to specific values with high posterior probabilities), the next step is to estimate posterior genotype probabilities at each SNP for all maternal individuals, as well as the paternal allele contributed to each offspring at each SNP. We first apply BORICE to simulated data, which demonstrates accurate estimation with surprisingly limited data (few SNPs and/or low coverage per SNP). The application to *Mimulus* illustrates individualized mating system estimation with real data and also how inferred maternal/paternal genotypes can be used as explanatory variables for spatial structure and to explain differences in mating system and fitness. The *Mimulus* data also shows the challenges that emerge when the volume and quality of sequence data varies greatly among samples.

2 | MATERIALS AND METHODS

2.1 | Theory

The posterior distributions for model parameters are estimated using the Metropolis–Hastings algorithm of Markov Chain Monte Carlo (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953). The first necessity is to calculate the likelihood of the data given model parameters. The movement of the chain through parameter space is governed mainly by how parameter changes affect the likelihood of the data. The final product of the analysis, the posterior distribution (or density) for each parameter, is estimated from the chain history (Hastings, 1970). For discrete parameters, the posterior probability for a particular value (say δ = 0 for offspring 1 of family 1) is the fraction of steps accrued in that particular

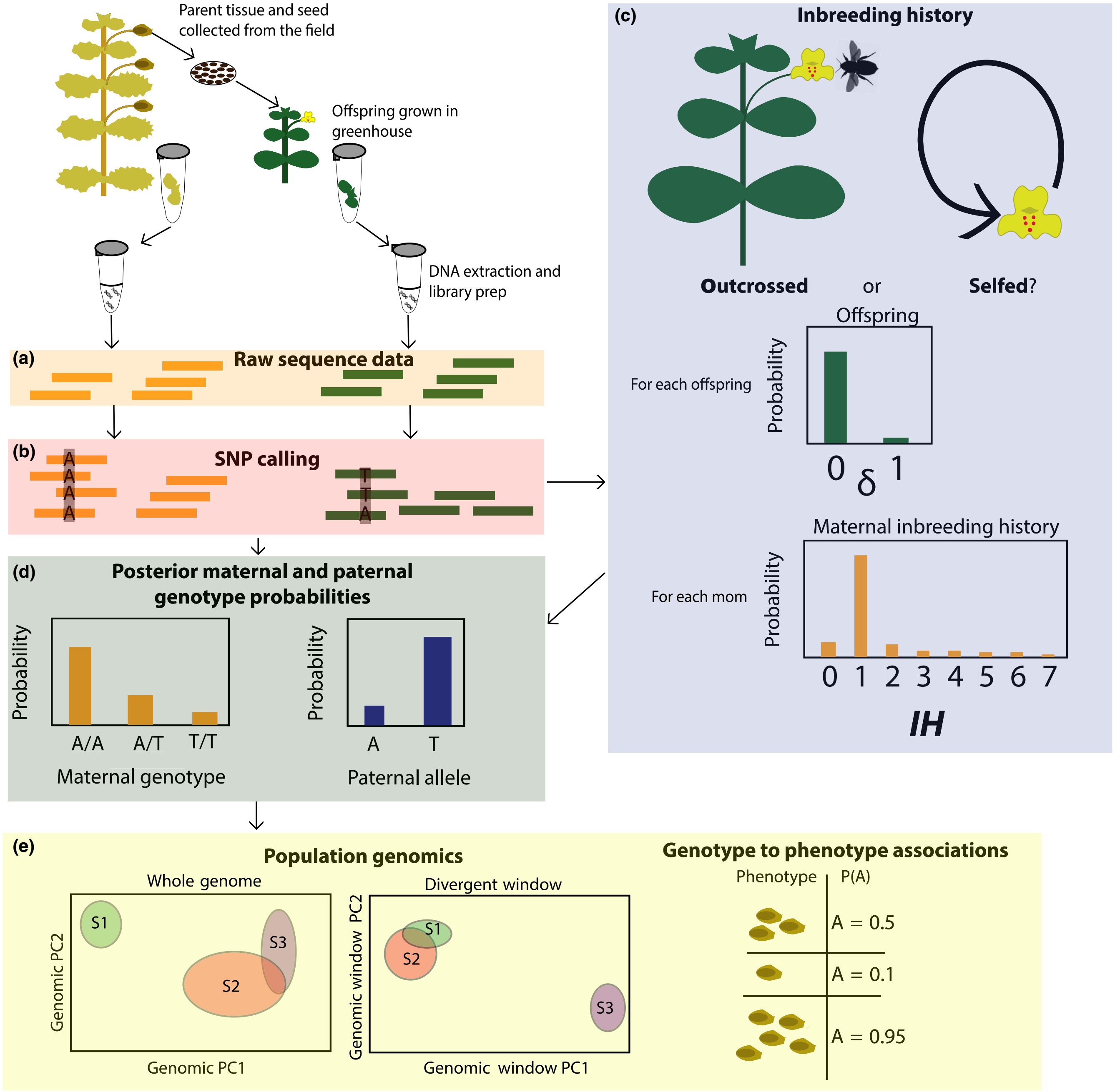

FIGURE 1 An overview of the analysis pipeline. Raw sequencing data (a) is mapped and filtered to SNPs within homologous loci (b). The first BORICE application uses these SNPs to classify (c) the probability that each offspring was generated from a selfing event and the inbreeding history for each maternal individual. The second BORICE application utilizes mating system estimates (c) with the SNP data (b) to derive posterior probabilities of (d) maternal genotypes and paternal alleles. These values can be used in down-stream applications (e). These include using genomic PCs to identify genome wide patterns of divergence between different sub-populations (S1, S2, S3) and window-based analyses to distinguish regions of the genome that display different patterns of differentiation. Finally, allele frequency differences between different phenotypic classes provide a preliminary mapping of phenotypic or fitness loci [Colour figure can be viewed at wileyonlinelibrary.com]

value. For continuous parameters, such as allele frequency, probability can be estimated from the chain visits within a range of values.

The model involves both hyperparameters and latent parameters which differ in how they enter the likelihood calculations and how updates are made. The three sets of hyperparameters are the vector of Q values (the ancestral allele frequency of each SNP – see description below), the vector of ζ values (the divergence parameters for each subpopulation), and t (the overall outcrossing rate). The latent parameters are the subpopulation specific allele frequencies (q), δ values (one for each offspring) and IH values (one for each parent). Within a population, the likelihood equations for parent/offspring genotype data at a single genetic locus in a mixed mating population (outcrossing and selfing) have been published previously (Fyfe & Bailey, 1951; Ritland & Jain, 1981), and implemented in both maximum likelihood (Ritland, 2002) and Bayesian inference frameworks (Koelling et al., 2012). Here, we extend the Bayesian model to biallelic SNPs with multiple subpopulations in the data set. Subpopulations are defined such that outcross matings are random within these units, but they may differ in allele frequency at any or all SNPs.

The log-likelihood (*LnL*) for the entire data set at a given SNP can be written:

$$LnL = \sum_{i=1}^{K} LnL_i \tag{1}$$

where *K* is the number of families and LnL_i is the log-likelihood of family *i* (all progeny of maternal plant *i*). Let δ_{ij} indicate outcrossed/self of the *j*'th progeny in family *i*, IH_i be the IH value (0–7) of maternal plant *i*, and q_i denote allele frequency in the subpopulation of this family. With these conventions, LnL_i can be written as a function of the genetic data at this SNP within the family:

$$LnL_i = Ln\left\{\sum_{m=0}^{2} P\left[MG=m|q_i,IH_i\right]\Pi_{j=1}^{n_i}P[O_{ij}|q_i,\delta_{ij},MG=m]\right\} \tag{2}$$

where MG is the maternal genotype (0, 1, and 2 are the reference homozygote, heterozygote, and alternative homozygote, respectively). $P\ [MG = m|q_i,\ IH_i]$ is the probability that MG = m given the maternal sequence data at this SNP (if it exists) conditional on q_i and IH_i. The product within Equation 2 is taken over all n_i offspring in the family. O_{ij} is the SNP data, however complete, for the *j*'th offspring of maternal individual i. $P[O_{ij}|q_i,\ \delta_{ij},\ MG = m]$ is the probability of O_{ij} given the maternal genotype, allele frequency, and whether this individual is outcrossed (δ_{ij} = 0) or selfed (δ_{ij} = 1).

BORICE can be run both with and without maternal genotype data. When there is no maternal genotype data, as is often the case in mating system studies, then

$$\begin{aligned}
P\left[MG=0|q_i,IH_i\right] &= q_i^2+f_iq_i\left(1-q_i\right)\\
P\left[MG=1|q_i,IH_i\right] &= 2\left(1-f_i\right)q_i\left(1-q_i\right)\\
P\left[MG=2|q_i,IH_i\right] &= \left(1-q_i\right)^2+f_iq_i\left(1-q_i\right)
\end{aligned} \tag{3}$$

where

$$f_i = 1-\left(\frac{1}{2}\right)^{IH_i}$$

When there is maternal genotype data, Equation 3 provide the priors (given f_i and q_i) that are multiplied by the likelihood of the observed maternal data (sequence reads) to give $P[MG = m|q_i,\ IH_i]$ for m = 0, 1, or 2. If the maternal genotype is known with certainty, only one of these terms will be nonzero.

Equation 2 also simplifies if offspring genotypes are known without error (see equation 1 of Koelling et al., 2012). With uncertainty, it is necessary to sum over all possible offspring genotypes weighted by their respectively likelihoods. For outbred progeny:

$$\begin{aligned}
P[O_{ij}|q,\delta_{ij}=0,MG=m] &= \Sigma_{(g=0)}^{2}\,\Sigma_{(x=0)}^{1}\,q^{(1-x)}(1-q)^{x}P[O_{ij}|MG=m,PA=x,OG=g]\\
&= \Sigma_{g=0}^{2}\Sigma_{x=0}^{1}q^{1-x}(1-q)^{x}M_{m,x\to g}U_{g\to O_{ij}}
\end{aligned} \tag{4}$$

Here, OG indicates the true offspring genotype (0, 1, or 2). PA is the paternal allele: 1 if the sire contributes the alternative base and 0 if he gives the reference base. $M_{m,x\to g}$ is the Mendelian probability that maternal genotype m crossed to paternal allele x produces offspring genotype g. $U_{g\to O_{ij}}$ is the probability that genotype g produces the observed SNP data (O_{ij}). For selfed progeny:

$$P\left[O_{ij}|q,\delta_{ij}=1,MG=m\right] = \Sigma_{g=0}^{2}S_{m\to g}U_{g\to O_{ij}} \tag{5}$$

where $S_{m\to g}$ is the Mendelian probability that selfing of maternal genotype m produces offspring genotype g.

When genotypes are known, $U_{g\to O_{ij}} = 1$ for the evident g and zero otherwise. With low-coverage genomic data, however, some individuals will have a single sequence read at a SNP. Even with relatively high coverage of SNPs, genotype calls may be less than certain owing to a diversity of possible errors (Mastretta-Yanes et al., 2015). SNP-calling algorithms, e.g., GATK (McKenna et al., 2010), routinely report GL scores (genotype likelihoods); three numbers indicating the likelihood of the data given each of the three possibilities for the true genotype (RR, RA, or AA). As described in Appendix S2, BORICE accepts two different input format for genomic data. The first takes GL scores literally, as $U_{g\to O_{ij}}$ for g = 0, 1, and 2, respectively. The second allows genotype quality, and thus the $U_{g\to O_{ij}}$ mapping, to vary among samples. Data format 2 requires an additional input file specify the number of loci called per individual.

The log-likelihood for the entire data set is calculated as a sum of Equation 1 applied to all SNPs. This treats all loci as independent, which is not likely true for linked SNPs. In this regard, BORICE neglects an important source of information, that selfing produces genomic tracts of Identity by Descent. With dense genotyping of SNPs located to chromosomes, these tracts can be identified (all SNPs are homozygous within an IBD region). Unfortunately, positional information is not available with anonymous markers (such as de novo assembled RADtags). Here, we thin the data to eliminate closely linked SNPs, choosing the single most informative SNP per RADtag (see Section 2.5). BORICE also assumes that all outcrossing is random (within each subpopulation) and that the paternity of outcrossed seeds within a family are determined independently. The latter assumption is relaxed when we consider likelihood of different family types (sires with maternal families).

2.1.1 | Population structure

The likelihoods (Equations 1–5) depend on allele frequency (q_i) in the subpopulation of family *i*. To allow differences among subpopulations, we implement the population structure model of Nicholson et al. (2002) which has been used most extensively as a component to the program Structure (Falush, Stephens, & Pritchard, 2003). The model considers each subpopulation as a derivative of an ancestral population with allele frequency Q. The extent of divergence, on average, is governed by a subpopulation specific divergence parameter (here denoted ζ). The ancestral population is not observable in the data, but Q_x should generally be within the range of q_{kx} values, where k indexes subpopulations and x indexes SNPs. Across all SNPs, the extent that q_{kx} deviates Q_x is determined by ζ_k, with $\zeta_k = 0$ meaning no divergence.

2.2 | MCMC configuration

The acceptance ratio is the product of the likelihood ratio, the prior ratio, and the Hastings' ratio. Our proposal mechanisms ensure that the Hastings' ratios are always 1, but prior ratios depend on the parameter being updated. The hyperparameters (the vectors of Q and ζ values, and t) each have a fixed prior distribution and likelihoods are a function of the latent variables. For t, we used a simple window proposal (a uniform interval around the current value). The prior is uniform (0, 1) and thus the prior ratio is 1. The likelihood is $t^{c_1}(1-t)^{c_2}$, where c_1 is the number of offspring with $\delta_{ij} = 0$ and c_2 is the number of offspring with $\delta_{ij} = 1$. These counts refer to the δ_{ij} values at the current point in the chain – as they change, so do likelihoods for updates to t.

For Q, the ancestral allele frequency at each SNP, we also use window proposals. The prior is uniform [0, 1] and prior ratio always 1. The likelihood is a function of the set of q_k, the allele frequencies within each subpopulation (k) which are treated as latent parameters (Nicholson et al., 2002):

$$L = \prod_k \beta\left[q_k, \frac{Q(1-\zeta_k)}{\zeta_k}, \frac{(1-Q)(1-\zeta_k)}{\zeta_k}\right] \qquad (6)$$

where $\beta[*]$ is the beta probability density function. For the divergence parameters, ζ_k, we use window proposals but the prior is beta distributed with user defined values for the two parameters (a,b). The prior ratio is $\frac{\beta[\zeta_k', a, b]}{\beta[\zeta_k, a, b]}$, where ζ_k' is the proposed value. The likelihood is the product across all SNPs:

$$L = \prod_x \beta\left[q_{kx}, \frac{Q_x(1-\zeta_k)}{\zeta_k}, \frac{(1-Q_x)(1-\zeta_k)}{\zeta_k}\right] \qquad (7)$$

where x indexes individual SNPs. For each of the window proposals, all proposals outside of the feasible range are rejected. The user specifies the width of window for each parameter.

Genetic data for maternal individuals and their offspring determine the likelihoods for the latent parameters (Equations 1–5), but the prior ratios are dependent on the hyperparameters. The likelihood for δ_{ij} updates is a product across all SNPs of family i (Equation 2). Updates to δ_{ij} are tested for each offspring individually in each iteration of the chain. The proposal is always to the alternative state (to 1 if δ_{ij} currently equals 0 and vice versa). The prior ratio is $t/(1-t)$ if the proposal is → outcrossed, and $(1-t)/t$ if the proposal is → selfed. A family level product across SNPs evaluates updates to IH_i values. The likelihood for updates to q_{kx} is a product over all families within subpopulation k for the specific SNP x. For proposed updates to IH, we sample a value from the prior – a value for IH (from 0 to 7) is chosen with probability $t(1-t)^{IH}$. The geometric distribution for IH is predicted at mating system equilibrium if there is no inbreeding depression. The prior ratio is not needed when updates are sampled from the prior distribution.

An iteration of the chain progresses through proposed updates in the following order: (a) ζ, (b) t, (c) δ_{ij} and IH_i values (progressing sequentially through families), (d) allele frequencies (ancestral and subpopulation specific for each SNP), and (e) the genotyping parameter (if using data format 2; see Appendix S1). BORICE includes two alternative options for the priors of δ_{ij} and IH_i (the default case described above is specified in the program control file as IHPriorModel = 0, deltaPriorModel = 0). Setting IHPriorModel = 1 decouples the IH prior from the current value of t in the chain. Instead of $P[IH = k] = t(1-t)^k$ (IHPriorModel = 0), $P[IH = k] = \frac{1}{2+3k+k^2}$; an equation obtained by integrating the geometric distribution over a uniform density for t. Users may choose IHPriorModel = 1 because inbreeding depression affects the distribution, elevating $IH = 0$ relative to higher categories. With alternative 2, setting deltaPriorModel = 1, specifies a uniform prior on δ (outcrossing and selfing equally likely a priori). This option, which decouples updates for specific progeny from t, could be useful if the outcrossing rate varies among subpopulations. The deltaPriorModel specification should not usually have a strong effect on inference: genomic data should speak strongly to whether an offspring is outcrossed or selfed, and as a consequence, the likelihood ratio will dominate the prior ratio.

The user has the option to collect posterior probabilities for maternal genotypes and paternal alleles (contributed to each outbred offspring) at each SNP over the course of an MCMC chain. At any step in the chain, we calculate posterior maternal genotype probabilities contingent on the sequence data (all individuals in the family) and the current values for q_{kx}, δ_{ij} and IH_i values (the latent parameters). The posterior for paternal allele (0 or 1) is likewise calculated given the maternal and specific offspring sequence data and the latent parameter values. The posteriors in a single step are added to a running sum (for each combination of SNP and parent) given that the numbers change in each step owing to updates in the latent parameters.

2.3 | Analysis pipeline

The initial step, processing of sequencing data (Figure 1, upper left), produces the BORICE formatted input files. The bioinformatic pipeline for our *Mimulus* application is described below, but alternative schemes will be necessary in other circumstances. For example, de novo construction of markers is needed when there is no reference genome (Catchen, Hohenlohe, Bassham, Amores, & Cresko, 2013; Eaton, 2014). We recommend a two-stage strategy to implement BORICE. In the first step, we infer offspring type

TABLE 1 Rates of correct, incorrect, and ambiguous classification of offspring as outbred or selfed in simulated data sets, with and without maternal genotype data. We considered cases with all offspring outcrossed (offspring type = outbred), all selfed, and half of each (mixed). Read depth is per SNP. Maternal plants were outbred in all cases

Details of simulation			Classification with maternal data			Classification without maternal data		
Number of SNPs	Read depth	offspring type	Correct	Ambiguous	Incorrect	Correct	Ambiguous	Incorrect
100	1	Outbred	0.733	0.260	0.007	0.463	0.534	0.003
100	1	Mixed	0.826	0.170	0.004	0.619	0.374	0.008
100	1	Selfed	0.929	0.068	0.003	0.823	0.168	0.009
100	10	Outbred	1	0	0	0.993	0.007	0
100	10	Mixed	1	0	0	0.999	0.001	0
100	10	Selfed	1	0	0	1	0	0
500	1	Outbred	0.999	0.001	0	0.948	0.050	0.002
500	1	Mixed	1.000	0.000	0	0.998	0.002	0
500	1	Selfed	1	0	0	1	0	0

and maternal inbreeding history from a subset of high quality (HQ) SNPs, delineated from high coverage across individuals. The HQ SNP set is used for the numerically intensive determination of δ_{ij} and IH_i values. Once δ_{ij} and IH_i are determined, we hold δ_{ij} and IH_i constant for the estimation of genotype probabilities of the complete SNP set. The maternal genotype and paternal allele probabilities for the full SNP set are the inputs for downstream analyses such as spatial population genetics and genotype-to-phenotype association.

2.4 | Simulation study

To evaluate the method under known conditions, we simulated genomic data sets and applied BORICE to estimate mating system parameters (Table 1). Each simulated data set consisted of 100 maternal families, each with four offspring. We considered (a) 100 SNPs sequenced at one read per SNP per individual (low coverage), (b) 100 SNPs sequenced at with 10 reads per SNP per individual (high coverage), and (c) 500 SNPs sequenced at 1 read per SNP (see Appendix S1 for additional details). For each of these cases, we applied BORICE with and without maternal data (maternal depth/coverage same as progeny) and tallied the number of offspring whose mating system was correctly estimated (posterior probability ≥0.95), incorrectly estimated, and ambiguously estimated.

BORICE correctly distinguishes outcrossed from selfed progeny with very limited data in terms of number of SNPs and coverage per SNP (Table 1). These patterns hold up when maternal plants are fully outbred (Table 1) as well as when plants are fully inbred. A range of cases are reported in Appendix S1. Briefly, these simulations show that maternal data improves classification in the most data-poor conditions. Inference of maternal IH is usually as precise as offspring outcross/self if there is maternal data, but can fail if direct genotyping data is absent. The choice of prior distribution for IH can matter in this situation. Finally, the simulations also show that intermediate frequency polymorphism are more informative than rare alleles. Larger numbers of SNPs are required if the population contains only rare alleles. In general, the simulation results are very encouraging given that genomic data sets will typically have much higher average numbers of SNPs (»500) and coverage per SNP (>1) than in Table 1. However, the simulations also indicate that uncertain outcomes are likely for the subset of individuals with low sequencing coverage.

2.5 | Experimental methods

2.5.1 | Field experiment

Mimulus guttatus (Phrymaceae, syn. *Erythranthe guttata*) is a western North American wildflower reliant on bees for pollination. The Quarry population is located in Oregon, USA (44.3454243 N, −122.1362023 W; Elevation ~1,200 m) and was likely colonized by *M. guttatus* only 30–40 years ago (Monnahan, Colicchio, & Kelly, 2015). We established transects across the various sections of the population in 2014 and observed developmental progress weekly through the summer of 2015. At the end of the growing season (23–24 July 2015), plants had completely senesced, but remained largely intact and holding whole seed pods. Subpopulations were identified as contiguous groups of plants isolated from other subpopulations by at least 3 m of open space. We randomly selected plants along transects within each subpopulation, excluding plants with fewer than three flowering nodes. We collected dried meristem tissue from each maternal plant and the earliest and latest main stem fruit (based on their locations in the inflorescence) and measured each plant for stem width, plant height, and numbers of branches, nodes, and flowers. The difference in anthesis between early and late flowers probably ranged from 2 weeks to 2 months, but in all cases, pollen from later nodes was deposited later in the season than pollen on earlier nodes of the same plant. Subpopulation 1 contained many more individuals than subpopulations 2 and 3, both of which were significantly larger than subpopulations 4 and 5 (Figure 2). Our sample sizes reflect the number of plants in each subpopulation.

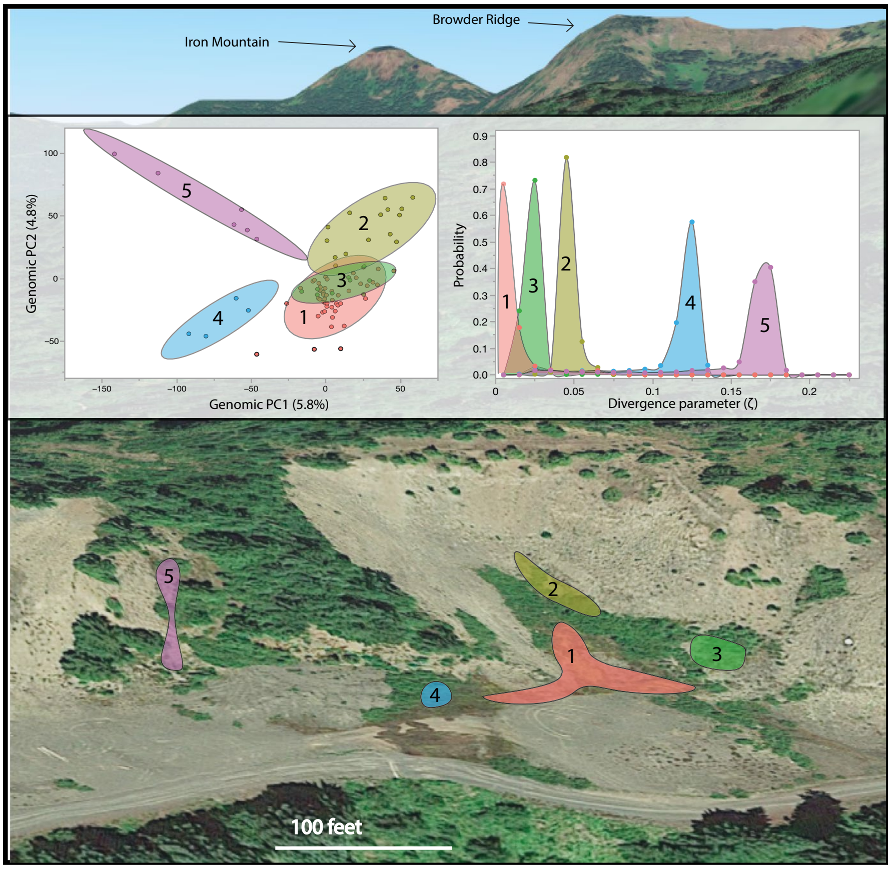

FIGURE 2 A map of Quarry sub-populations delineated by position. Left inset: the scoring of individual plants for genomic principal components. Right inset: The posterior density function estimates for ζ (sub-population divergence) values [Colour figure can be viewed at wileyonlinelibrary.com]

2.5.2 | Progeny grow-up, sequencing and genotyping

We grew up to five plants from each field collected fruit to flowering for meristem tissue collection. We extracted DNA from field collected and greenhouse grown tissue using a CTAB-buffer protocol optimized for Mimulus (Holeski et al., 2014). We constructed multiplexed shotgun genotyping (MSG; Andolfatto et al., 2011) libraries for sequencing, using restriction enzyme Csp6I and size selected for fragments 248–302 bp in length. These libraries were sequenced across both lanes of an Illumina HiSeq 2500 Rapid Run paired-end 100 bp. We obtained sequences from 349 plants (parents and offspring combined), although with great variance in coverage among samples.

To each sequence read pair, we applied Scythe (https://github.com/vsbuffalo/scythe/) to remove adaptor contamination and then Sickle (https://github.com/najoshi/sickle/) to trim low quality sequence. We used BWA-mem (Li, 2013) to map read pairs to the V2 genome build of *M. guttatus* (https://genome.jgi.doe.gov/portal/pages/dynamicOrganismDownload.jsf?organism=Mguttatus). We piped the output from Samtools mpileup (Li et al., 2009) to the Varscan v2.3.6 function mpileup2snp to call SNPs (Koboldt et al., 2009). From the resulting VCF file, we eliminated SNPs not called in at least 50 plants and identified the single most informative SNP per RAD-tag. Among SNPs with a mapping quality score ≥30, this SNP had the highest count of minor homozygote plus heterozygotes. We eliminated loci with non-Mendelian inheritance or excessively high or low coverage and formatted the remaining 30,318 SNPs for input to BORICE.

We identified 4,577 SNPs as high quality (HQ set) based on coverage; each of which was scored in at least 180 plants. We first ran BORICE on the HQ set with ChainLength = 2,000, burnin = 1,000, and a thinning frequency of 2. We aggregated results from 10 independent chains to calculated posterior probabilities. After compiling results from the HQ runs, we fixed δ_{ij} and IH_i of offspring and maternal parents according to the posterior probabilities from the HQ runs. We then ran BORICE on the larger SNP set, one chromosome at a time, to estimate posterior probabilities of maternal genotypes and paternal alleles at each SNP. Finally, given the genotype posterior probabilities, we calculated the likelihoods for each family for all possible siring patterns among outbred progeny.

2.5.3 | Genomic PC analyses and contrast to IM sequenced lines

For downstream analyses comparing genomic patterns across individuals, we distilled SNP posterior genotypes probabilities from both male and female parents using principle components of genotype scores at each SNP. We separated maternal genotype scores and paternal allele scores into separate analyses. For diploid maternal genotypes, the score is 2*(posterior probability of RR) plus the posterior probability of RA. For the paternal allele, the score is the posterior probability that the pollen grain carries the reference base. We filtered out individuals with data at <5,000 SNPs in the full SNP set,

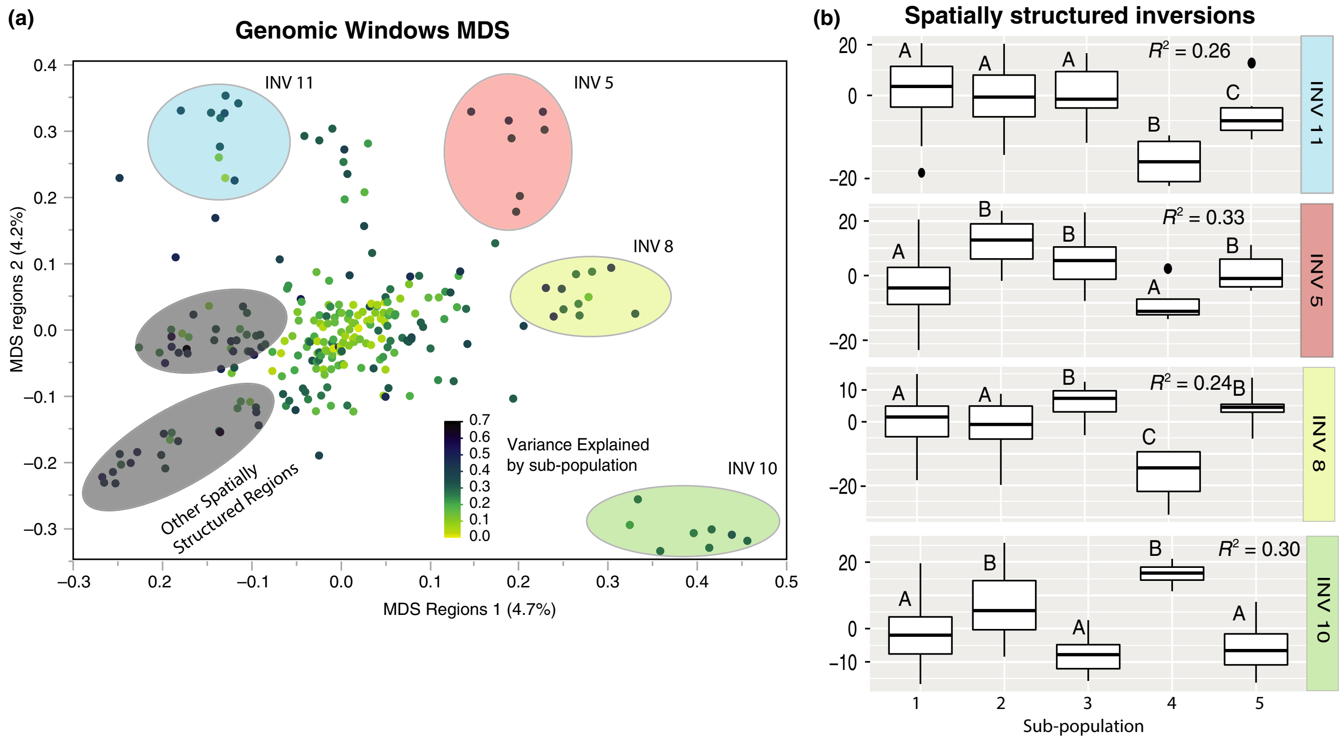

FIGURE 3 (a) Genomic PCs in MDS space. Points represent PC1 of each of the 263 genomic windows. Four of the outlier groups corresponded to known inverted regions (see four colored bubbles). Coloring points by the proportion of the variance between individuals that can be explained by their sub-population shows that these outlier windows tend to show elevated spatial structure relative to the remainder of the genome. (b) Box-plots showing subpopulation differences in structural variant PC1 values. R2 is the proportion of variance that can be explained by subpopulation, and letters signify significant differences ($p < 0.01$) identified through post-hoc tests [Colour figure can be viewed at wileyonlinelibrary.com]

and all subsequent analyses were done separately on the remaining 83 maternal plants and 213 (inferred) pollen grains. We calculated allele frequencies within the pollen pool from the pollen genotype scores at each SNP, parsing estimates by both subpopulation and by early/late fruit. Estimates from maternal plants and pollen grains were compared to allele frequencies at the nearby well-studied Iron Mountain (IM) population of *M. guttatus*. We identified 14,079 SNPs in our full Quarry data set that were also ascertained in the IM SNP set reported by Troth, Puzey, Kim, Willis, and Kelly (2018).

We calculated genome-wide principal components (including all SNPs) using the *nipals* method (pcaMethods 1.75.1) to account for missing data and unit variance scaling (variance percentages reported in Appendix S2). Next, we used lostruct (Li & Ralph, 2018) to calculate genomic window PCs across the genome, and saved the first two principal components, to examine variation in relatedness across the genome. We ran this separately for each of the 14 chromosomes specifying 100 SNPs per window. This produced 13–30 windows per chromosome ranging in physical size from 380 kb to 4 mb. For chromosomes 8 and 12, we re-ran the program stipulating 120 SNPs per window to avoid very small windows at chromosome ends (coordinates and variance percentages are reported in Appendix S3). We performed post-hoc multidimensional scaling (MDS) analysis to summarize variation across individual PC scores within a given window. When plotted in two-dimensions, all genomic windows and found that those within known structural variants in *M. guttatus*, the inversions on chromosomes 5, 8, and 10, as well as the meiotic drive locus on 11 (Flagel et al., 2019; Monnahan & Kelly, 2017), are outliers (Figure 3). We combined windows within each structural variant region and calculated PCs specific to each of these four loci (including 507–745 polymorphisms per locus). The three sorts of PC (genome-wide, structural variant, and window) are used to investigate spatial structure and as predictors of maternal plant phenotypes, fitness measurements, and individual selfing rates.

2.5.4 | Genetic effects on phenotype

To identify genetic associations with plant phenotype (log transformed flower count, and arcsin transformed percent selfing), we fit a general linear mixed model. The fixed effects were subpopulation (to control for environmental differences among populations), the inversion specific PC1 scores, and the first 3 PCs of the remaining genomic background. The random effects were the unexplained genetic variants (not caused by genetic fixed effects) and the environmental deviation. We calculated a relatedness matrix to predict the covariance among plants by adapting equation 3 of Yang, Lee, Goddard, and Visscher (2011) to our genotypic scores (Appendix S4). The PCs and additive variance are complementary methods to characterize the genomic background

(Bradbury et al., 2007; Yu et al., 2006). We performed these calculations using mcmcGRM (Hadfield, 2010) with default priors for both the residual and additive variances. We performed 13,000 total iterations and discarded the first 3,000 as burnin. To scan for specific genomic windows associated with phenotype, we first calculated residuals for each phenotype given the significant fixed effect identified in the model just described, and then used the response screening tool in JMP (FDR adjusted *p*-value < 0.05, JMP Pro v14.0 Sas Institute Inc.) to identify other genomic windows associated with phenotypic variation. Terms identified as significant (false discovery rate <5%) were added to the full model (including subpopulation ID, significant genomic PCs and INV PCs) to predict phenotype in mcmcGRM. In order to look for individual SNPs driving phenotype associations, we once again used the response screening approach to identify significant SNPs (FDR adjusted *p*-value <.05, JMP Pro v14.0 SAS Institute Inc.), followed by evaluation in the full model MCMC.

3 | RESULTS

3.1 | Inference of mating type and inbreeding history

A total of 235 of the 250 progeny were genotyped at >500 SNPs in the HQ set. Of these, 208 (88.5%) are called as outcrossed with high confidence (posterior probability >0.99 that δ = 0), 23 (9.8%) are called as selfed (posterior probability >0.95), and four have intermediate outcomes (see Appendix S5 for detailed output). Selfed progeny are well dispersed across maternal plants. Only one maternal plant (109) has more selfed progeny than outcrossed (3 vs. 0). There was no difference in estimated selfing rate between early and late flowers, 13% vs. 9% ($X^2_{[1]}$ = 0.84, *p* = .36). The 95% credibility interval for the overall outcrossing rate is 0.8–0.9.

Considering maternal inbreeding histories, 62 maternal plants are confidently called as fully outbred, seven are confidently called as first generation selfs, and four are called as more highly inbred (IH ≥ 2). The remainder (42 of 115) are ambiguous, usually because these plants lacked maternal genotype data and had weak progeny data (Appendix S6 for detailed output). A total of 57 families had two or more outcrossed progeny allowing us to assess diversity in paternal genotypes. While the BORICE model assumes that all progeny were sired independently (or selfed), two progeny sired by the same individual will have more similar pollen genotypes than those sired by different individuals. Using the pollen allele posterior probabilities, we consider the likelihood of all possible siring configurations among the outbred progeny of each family (Appendix S7). This analysis indicates the progeny were entirely full sibs for 15 families (26%). In 12 families (21%), all progeny were half sibs, each sired by a distinct plant. The remaining families were mixtures of half and full sibs. Across the 191 outbred progeny in the entire data set, the average probability that a sib was sired by the same plant was 43%. There was a significantly higher probability that two individuals from the same fruit shared a sire (54%) than individuals from separate fruits (33%; X^2 = 11.82, *p* = .0006).

3.2 | Spatial structure of maternal plants inferred from progeny genotypes

The posterior density functions for subpopulation differentiation, ζ (Figure 2, right inset), suggest subpopulations 1–3 are minimally differentiated, subpopulations 4 and 5 much more so. These patterns are corroborated by genome-wide principle component scores calculated for maternal plants, which exhibit highly significant differentiation among subpopulations for the first 3 PCs (*p* < .0001 for each).

These explain 5.8% (4.2%), 4.8% (3.9%), and 4.5% (2.7%) of the genome wide variance in maternal (male) scores, respectively. However, the window PCs indicate that this differentiation varies greatly over the genome. The percent of variance in window PC1 that could be explained by subpopulation varied dramatically from 1.7% to 66.8%, suggesting that while some genomic regions are homogeneous across Quarry, others show extreme spatial structure. Multidimensional scaling allowed us to visually identify genome windows that are highly differentiated across individuals as outliers in MDS space (Figure 3a). About 50% of the genome shows very little spatial structure (points central in MDS space). However, windows located within the structural variant regions show as clear outliers with elevated divergence across the Quarry. The ovals enclosing inversions (Figure 3a) are drawn using the approximate breakpoints for each polymorphism as reported in Monnahan and Kelly (2017).

The first PC explained an average of 25% of the variance for windows outside structural variants, but this increased to 35% within these regions (Figure 3b). Not only was a higher proportion of variance explained by spatial structure within these windows, but in the case of INV5 and INV10, the spatial structure within these regions was very distinct from other spatially structured windows. For instance, of the nine windows that showed the strongest evidence (R^2 > .235) of subpopulations 1 and 4 being divergent from the remainder of the subpopulations, seven were the windows within INV5. Likewise, INV10 was very similar across subpopulations 1, 3, and 5; which is atypical of the remaining genome. INV8 and MDL11 both showed substantial spatial structure, but the patterns of spatial segregation within these inversions were not substantially different from other structured windows (Figure 3a).

Across windows, there is a highly significant (*p* < .0001) positive correlation between the percent of variation explained by subpopulation for male and female parents (*r* = .8). The five most common patterns of spatial variation across the first PC of the 263 genomic windows were as follows: No significant spatial effect (47.5%); regions where subpopulation 4 was highly divergent from all other subpopulations (11.8%); regions where subpopulation 5 was divergent (18.2%); regions where 4 and 5 were divergent from 1, 2, and 3 (15.9%), and regions where 2, 3, and 5 were divergent from 1 and 4 (6.1%). Any genomic region where <15% of the variation could be explained by subpopulation was grouped in the first category. This tabulation indicates that for about half of the genome there is little differentiation between the different Quarry subpopulations.

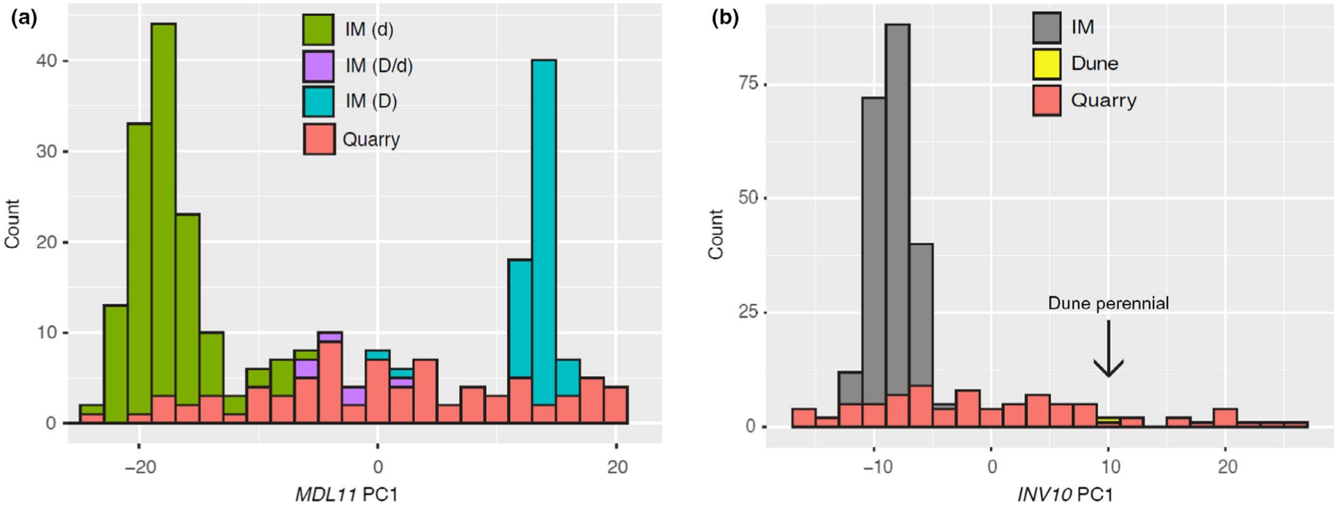

FIGURE 4 The distribution of PC1 values within the MDL11 (A) and INV10 (B) genomic regions are reported for IM lines, Quarry maternal plants, and the DUN genome sequence [Colour figure can be viewed at wileyonlinelibrary.com]

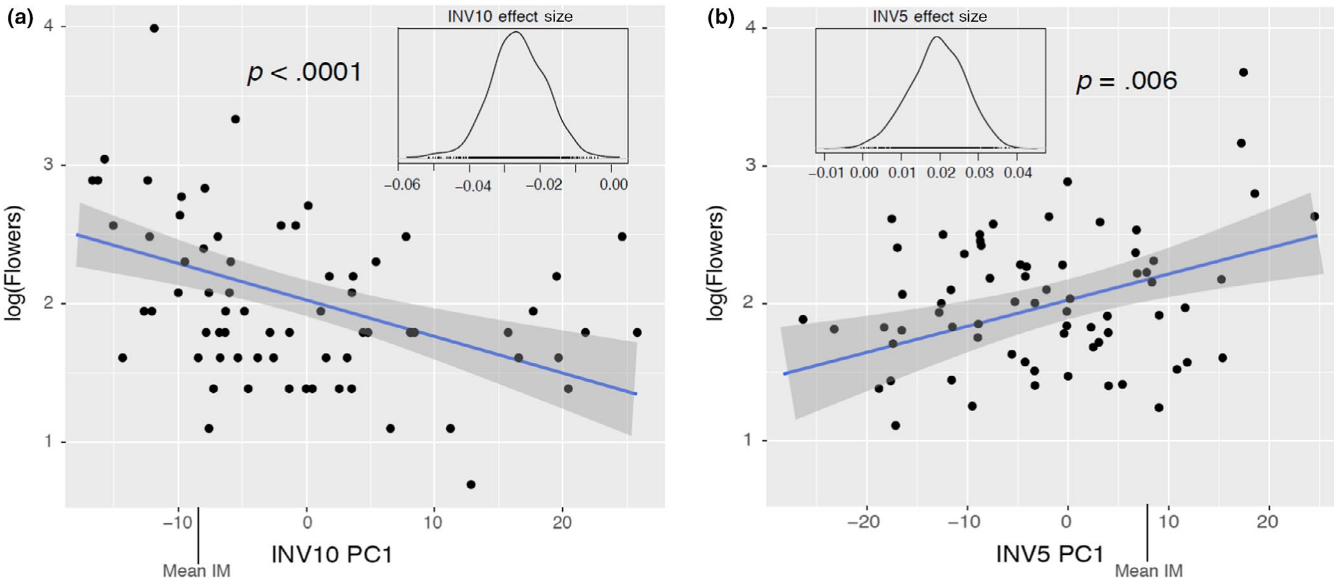

FIGURE 5 The estimated effects of INV10 (A) and INV5 (B) on female fecundity. The insets are the posterior density functions for effect. The Iron Mountain inversion orientations have low values for INV10 PC1 but high values for INV5 PC1 [Colour figure can be viewed at wileyonlinelibrary.com]

3.3 | Genomic PC scores effectively diagnose Inversion types

To confirm that window PCs diagnose structural polymorphisms, we applied the PC scoring formulae calculated from the Quarry plants to each of the sequenced lines from Iron Mountain (IM). The IM lines were previously scored at the Meiotic Drive Locus (MDL11) by Troth et al. (2018) and these calls show a striking bimodal distribution for PC1 (Figure 4a). The few heterozygous IM lines (purple) are intermediate for PC1. Quarry plants (red) exhibit PC1 values that span the range of alternative homozygotes within IM. The other inversions (5, 8, 10) are not known to be polymorphic within IM. However, the IM orientation for INV10 is at one end of the spectrum, clearly distinct from DUN, a line that exhibits the alternative arrangement relative to the IM lines (Figure 4b). As with MDL 11, we see that the Quarry plants span the range established by fully genome sequenced plants with known inversion types.

3.4 | Pollen allele frequencies

The nearby IM population consists of rapidly flowering genotypes (Troth et al., 2018), and thus provides a natural comparison to the early and late allele frequencies in pollen at Quarry. The SNP base most frequent in IM tends to be elevated in early relative to late fruits: the regression of allele frequency difference between early and late fruits onto allele frequency in IM is highly significant

(p < .0001, F = 33.6, n = 14,179). Considering SNPs more than 4% divergent (in either direction) between early and late, change was highly correlated with IM allele frequency (p < .001, F = 42.2, n = 4,272, Appendix S8), while for the remainder of the SNPs there was no pattern. Next, we compared early and late pollen allele frequencies to subpopulation allele frequencies. We found that SNP variants at higher frequencies in the early pollen also tended to be more common in subpopulations 1 and 4, reduced in frequency in subpopulation 2. Subpopulation 2 is located between 20 and 60 feet above the main quarry site, while subpopulation 4 is in a particularly swampy area in the lower section of the quarry.

3.5 | Genome wide association mapping

Our fitness measure, lifetime flower production, was strongly affected by the inversion genotype on chromosomes 10 (p < .001) and 5 (p = .006, R^2 for linear regression = .20, Figure 5, Supporting Appendix S9). The IM orientation for each inversion was associated with increased fitness. The difference between inbred and outbred plants was not significant, although this test has low power owing to the small number of inbred maternal plants. For plant height, only inversion 10 had a significant effect, positive for the IM karyotype (Appendix). There was also a marginal effect of genomic PC3. For the fraction of offspring produced by selfing, two genomic windows emerged as highly significant predictors, window 69 (p < .0001, chromosome 4: 17,979,552–18,798,961) and 184 (p < .008, chromosome 11: 5,201,765–5,876,977, Supporting Appendix S9). R^2 = .28 for the linear regression.

4 | DISCUSSION

4.1 | BORICE can estimate mating system parameters at the individual level

Genomic sequencing allows individual-level estimation of mating system parameters and thus the extent to which mating system parameters vary within and among populations. Our simulations illustrate that low-level sequencing (a single sequence read at as few as 500 SNPs) can effectively diagnose offspring as outcrossed or selfed and determine whether specific maternal individuals are inbred and to what extent (Table 1, Appendix S1). Real data is invariably more error prone than simulated data. Despite this, over 98% of progeny genotyped at >500 SNPs were scored definitely as either outcrossed or selfed in our empirical *M. guttatus* data set. We did not detect any difference in selfing rate between flowers produced early vs. late in the growing season, but the experiment illustrates the diversity of downstream applications enabled by BORICE. Analyses of posterior probabilities for both mating system parameters and individual genotypes demonstrate that (a) inbred plants are reproductively successful in Quarry, (b) the population exhibits remarkable internal spatial structure strongly associated with inversion polymorphisms, and (c) a mating system survey can provide a preliminary mapping of both selfing rate and fitness determining loci.

The population outcrossing rate estimate is narrowly bounded (95% credibility interval 0.8–0.9) because error comes almost entirely from the finite number of offspring and not uncertainty about their status. Outbred progeny from a single fruit were often full siblings, which is unsurprising because siring may result from a single pollinator visit. However, shared male parentage across early and late fruits is notable given that these flowers would not have been open simultaneously. Most likely, the shared sire in these cases is a physically proximal plant, one that is routinely sampled by bees during the same pollinator foraging run (Thomson, Slatkin, & Thomson, 1997). Across the data set, siblings were more frequently half (different sires) than full (same sire). The fact that half-siblings often produced from the same fruit suggests frequent pollen carryover (Thomson & Plowright, 1980): visitors are delivering mixtures of pollen from multiple, previously visited, plants.

4.1.1 | Successful reproduction by inbred adults

In a survey of mixed-mating species, Scofield and Schultz (2006) noted that the mean adult inbreeding coefficient (F) was much lower in long-lived than short-lived plants. This suggests stronger inbreeding depression in the long-lived species. *M. guttatus* is short-lived so our finding that inbred individuals successfully survive to reproduce is not surprising, although inbreeding depression is very severe in some populations of this species (Carr & Dudash, 1997; Carr, Fenster, & Dudash, 1997; Willis, 1993). The salient result is that genome-wide genotyping effectively diagnoses individual parental plants as inbred or outbred. This is essential for testing whether any inbred progeny succeed in species where the adult F is close to zero. Admittedly, our inference of adult inbreeding history was less precise than for progeny outcross/self (Appendix S6). This occurred because our maternal plants senesced in the field, reducing tissue quality for DNA extraction. This will routinely prove an issue for annual plants, but much less for perennials where the question of inbred success is most in doubt.

4.2 | Quantitative and population genetics using genomic window principal components

In mating system estimation, genotyping first serves to establish whether offspring are outcrossed or selfed (δ parameters of BORICE) and the inbredness of parents (IH parameters). However, genome-wide marker data can be exploited to address a diversity of additional questions. The challenge here is that many SNPs will have low quality calls or are completely uncalled in many plants. A number of methods have been developed to predict geography, phenotype, and fitness from uncertain genotype data (e.g., Gompert et al., 2010; Monnahan et al., 2015; Parchman et al., 2012; Wessinger, Kelly, Jiang, Rausher, & Hileman, 2018). We here explore the use of Principle Component scores defined within genomic windows (Li & Ralph, 2018) to distill genotype posterior probabilities. It is essentially a method to aggregate signal from closely linked SNPs. For *Mimulus*, we find that window PCs effectively identify segregating inversion polymorphisms (Figures 3–5) without any extrinsic information regarding the genomic location or character of these loci.

Figure 5 suggests that, whenever possible, researchers should include phenotypic and/or fitness measurements of maternal individuals in mating system experiments. After controlling for population structure and genetic background, two of the inversion polymorphisms remained as significant predictors of phenotype and fitness measurements. Previous field experiments have demonstrated fitness effect for both the inversions on chromosome 6 (Lee, Fishman, Kelly, & Willis, 2016) and the Meiotic Drive Locus on chromosome 11 (Fishman & Kelly, 2015; Fishman & Saunders, 2008). This study provides the first evidence of such effects for the inversions on chromosomes 5 and 10. More generally, the combined spatial structure and fitness results hint at a potentially crucial role of structural variants in mediating adaptation in the face of gene flow on a small spatial scale.

A key motivation for individual estimation of mating system parameters is to identify the genetic, phenotypic, and ecological determinants of variation in selfing rate. Alhough we did not detect differing selfing rates between early and late flowers, the association mapping analysis identified a putative genetic modifier of selfing rate on chromosome 4. Specifically, a window PC defined from 18.0 mb to 18.8 mb of this chromosome is a significant predictor of selfed offspring per maternal plant. This window is a substantial length of DNA containing 140 genes. We tested individual SNPs within the RAD-tags for association with selfing, and while the strongest signals were located on chromosome 4, they were at SNPs just outside the implicated window. These results are intriguing, but entirely preliminary. Confirmation of this selfing rate QTL will require a follow-up field experiment, perhaps coupled with high quality genotyping of chromosome 4 markers. However, the chromosome 4 association serves as a useful example of how a genomic mating system experiment can screen for ecologically relevant loci.

4.2.1 | Finescale spatial structure

Estimation of mating system parameters sets the stage for population genetic analyses. The BORICE ζ statistics allow divergence among the subpopulations in allele frequency, which is essential to properly assess the probability that an offspring is outcrossed (how likely the pollen allele matches or mis-matches alleles from the maternal parent). Quarry subpopulations exhibit divergence (Figure 2), but the ζ statistics do not capture the varying pattern across the genome. The window PCs indicate that about half of the genome exhibits minimal spatial differentiation, but specific genomic regions are highly divergent, particularly the inversion regions (Figure 3). For example, divergence of subpopulation 4 is largely due to the inversion on chromosome 8 (Lowry & Willis, 2010) and the meiotic drive locus on chromosome 11 (Fishman & Saunders, 2008). In contrast, subpopulation 5 is divergent at SNPs dispersed across the genome, not specifically inversion regions. Previous studies had established that Quarry is a hybrid swarm of annual and perennial genotypes (Monnahan et al., 2015; Monnahan & Kelly, 2017), but not the internal spatial patterning of variation. The window PC analysis suggests that during the incipient stages of microhabitat sorting and local adaptation, subpopulations can be nearly indistinguishable for much of the genome but highly divergent at particular loci. Mimulus inversions have previously been implicated in local adaptation across large geographic regions (Lowry & Willis, 2010; Twyford & Friedman, 2015), but the present study suggests they might play a role in finescale local adaptation (Figure 2). Reciprocal transplant experiments are needed to test this hypothesis.

Maternal plants are sampled at a known location and thus the posterior probabilities associated with maternal genotypes are immediately amenable to spatial population genetic analyses. In contrast, the BORICE inferred pollen genotypes are from plants at unknown locations. However, we can compare allele frequencies inferred from pollen to those obtained from spatially located females, and to data from other populations. There was generally a strong correlation between male and female allele frequencies within subpopulations, which is not surprising given that localized mating is needed to maintain spatial structure. The pollen populations from early and late fruits were interestingly different. Male allele frequencies in early fruits were most similar to subpopulations 1 and 4 while late males were more similar to subpopulation 2. Subpopulation 2 is located above subpopulations 1 and 4 and partially blocked from early morning sun making it a later flowering region. However, it is possible these differences are more due to genetic effects on flowering time than the location of male parents. Early pollen allele frequencies are more similar to those in the Iron Mountain population (about 6 km distant) than is late season pollen (Appendix S8). Iron Mountain is not contributing alleles by direct migration to Quarry (excepting rare long-distance dispersal), but it is a rapid flowering population. This suggests an effect of polygenic variation affecting flowering time. More generally, the analysis of siring illustrates how population genomic analyses are fully dependent on the characterization of mating system. We cannot infer allele frequencies in the cross-pollen population without first determining whether offspring are outcrossed or selfed.

In summary, genomic sequencing methods have transformed experimental approaches in molecular ecology, particularly in spatial population genetics/phylogeography (Hodel et al., 2017; Jeffries et al., 2016; Ruegg, Anderson, Boone, Pouls, & Smith, 2014), as well as the direct study of natural selection (Chen et al., 2019; Flanagan & Jones, 2017; Monnahan et al., 2015; Soria-Carrasco et al., 2014). Despite limited progress thus far, genome-wide genotyping offers considerable promise to advance mating system investigations by addressing previously impenetrable questions. BORICE can be useful for this purpose, at least with regard to questions that require estimates specific to individual plants or animals. Our application of BORICE to the Quarry population of *M. guttatus* also suggests novel directions for mating system experiments, particularly if genotypes can be related to geography or phenotype or fitness. These applications may have limited power, but can provide a critical first step to identify ecologically important regions of the genome.

ACKNOWLEDGEMENTS

We thank C. Crawford, M. Warner, M. Orive, M.T. Holder, P. Stokes, A. Gaudinier, and M. Akman for input on the project and/or manuscript. We thank P. Vogler and J. Colicchio for field assistance and C. Friesen (U.S. Forest Service) for site access. This work was supported by grants from the National Institutes of Health to JKK (R01 GM073990) and the KU Botany Endowment to JC.

AUTHOR CONTRIBUTIONS

J.C. and P.J.M. conceived and performed the field experiment. J.K.K. wrote the theory and the revised BORICE program. C.A.W. performed the simulation study. K.B., and J.R.K. completed all wet laboratory work. J.C., and J.K.K. conducted data analysis. J.C., C.A.W., and J.K.K. wrote the manuscript with contributions from the other coauthors.

DATA AVAILABILITY STATEMENT

Python scripts used to perform analyses are contained in Supporting Information 1 and are hosted on Github (https://github.com/jkkelly/Borice.genomic). The publicly available version of BORICE is open source and written in Python 2.7 (http://www.python.org/). Sequence data is available at the NCBI Sequence Read Archive as study PRJNA544272.

ORCID

Carolyn A. Wessinger https://orcid.org/0000-0003-3687-2559
John K. Kelly https://orcid.org/0000-0001-9480-1252

REFERENCES

Andolfatto, P., Davison, D., Erezyilmaz, D., Hu, T. T., Mast, J., Sunayama-Morita, T., & Stern, D. L. (2011). Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Research*, *21*(4), 610–617. https://doi.org/10.1101/gr.115402.110

Baker, H. G. (1955). Self-compatibility and establishment after 'long-distance' dispersal. *Evolution*, *9*(3), 347–349.

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*(19), 2633–2635. https://doi.org/10.1093/bioinformatics/btm308

Campbell, R. B. (1986). The interdependence of mating structure and inbreeding depression. *Theoretical Population Biology*, *30*, 232–244. https://doi.org/10.1016/0040-5809(86)90035-3

Carr, D. E., & Dudash, M. R. (1997). The effects of five generations of enforced selfing on potential male and female function in Mimulus guttatus. *Evolution*, *51*, 1797–1807.

Carr, D. E., Fenster, C. B., & Dudash, M. R. (1997). The relationship between mating-system characters and inbreeding depression in Mimulus guttatus. *Evolution*, *51*(2), 363–372.

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: An analysis tool set for population genomics. *Molecular Ecology*, *22*(11), 3124–3140. https://doi.org/10.1111/mec.12354

Chen, N., Juric, I., Cosgrove, E. J., Bowman, R., Fitzpatrick, J. W., Schoech, S. J., … Coop, G. (2019). Allele frequency dynamics in a pedigreed natural population. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(6), 2158–2164. https://doi.org/10.1073/pnas.1813852116

Cockerham, C. C., & Weir, B. S. (1984). Covariances of relatives stemming from a population undergoing mixed self and random mating. *Biometrics*, *40*, 157–164. https://doi.org/10.2307/2530754

Duminil, J., Hardy, O. J., & Petit, R. J. (2009). Plant traits correlated with generation time directly affect inbreeding depression and mating system and indirectly genetic structure. *BMC Evolutionary Biology*, *9*(1), 1–14. https://doi.org/10.1186/1471-2148-9-177

Eaton, D. A. R. (2014). PyRAD: Assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics*, *30*(13), 1844–1849. https://doi.org/10.1093/bioinformatics/btu121

Eckert, C. G., Kalisz, S., Geber, M. A., Sargent, R., Elle, E., Cheptou, P.-O., … Winn, A. A. (2010). Plant mating systems in a changing world. *Trends in Ecology & Evolution*, *25*(1), 35–43. https://doi.org/10.1016/j.tree.2009.06.013

Ellis, T. J., Field, D. L., & Barton, N. H. (2018). Efficient inference of paternity and sibship inference given known maternity via hierarchical clustering. *Molecular Ecology Resources*, *18*(5), 988–999. https://doi.org/10.1111/1755-0998.12782

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, *164*(4), 1567–1587.

Fishman, L., & Kelly, J. K. (2015). Centromere-associated meiotic drive and female fitness variation in Mimulus. *Evolution*, *69*(5), 1208–1218. https://doi.org/10.1111/evo.12661

Fishman, L., & Saunders, A. (2008). Centromere-associated female meiotic drive entails male fitness costs in monkeyflowers. *Science*, *322*(5907), 1559–1562.

Flagel, L. E., Blackman, B. K., Fishman, L., Monnahan, P. J., Sweigart, A., & Kelly, J. K. (2019). GOOGA: A platform to synthesize mapping experiments and identify genomic structural diversity. *PLOS Computational Biology*, *15*(4), e1006949. https://doi.org/10.1371/journal.pcbi.1006949

Flanagan, S. P., & Jones, A. G. (2017). Genome-wide selection components analysis in a fish with male pregnancy. *Evolution*, *71*(4), 1096–1105. https://doi.org/10.1111/evo.13173

Flanagan, S. P., & Jones, A. G. (2019). The future of parentage analysis: From microsatellites to SNPs and beyond. *Molecular Ecology*, *28*(3), 544–567. https://doi.org/10.1111/mec.14988

Forrest, J. R. K. (2015). Plant–pollinator interactions and phenological change: What can we learn about climate impacts from experiments and observations? *Oikos*, *124*(1), 4–13. https://doi.org/10.1111/oik.01386

Fyfe, J. L., & Bailey, N. T. J. (1951). Plant breeding studies in leguminous forage crops I. Natural cross-breeding in winter beans. *The Journal of Agricultural Science*, *41*(4), 371–378. https://doi.org/10.1017/S0021859600049558

Gezon, Z. J., Inouye, D. W., & Irwin, R. E. (2016). Phenological change in a spring ephemeral: Implications for pollination and plant reproduction. *Global Change Biology*, *22*(5), 1779–1793. https://doi.org/10.1111/gcb.13209

Gompert, Z., Forister, M. L., Fordyce, J. A., Nice, C. C., Williamson, R. J., & Alex Buerkle, C. (2010). Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of Lycaeides butterflies. *Molecular Ecology*, *19*(12), 2455–2473. https://doi.org/10.1111/j.1365-294X.2010.04666.x

Goodwillie, C., Kalisz, S., & Eckert, C. G. (2005). The evolutionary enigma of mixed mating systems in plants: Occurrence, theoretical explanations, and empirical evidence. *Annual Review of Ecology, Evolution, and Systematics*, *36*(1), 47–79. https://doi.org/10.1146/annurev.ecolsys.36.091704.175539

Govindaraju, D. R. (1988). Mating systems and the oportunity for group selection in plants. *Evolutionary Trends in Plants*, *2*, 99–106.

Grossenbacher, D., Briscoe Runquist, R., Goldberg, E. E., & Brandvain, Y. (2015). Geographic range size is predicted by plant mating system. *Ecology Letters*, *18*(7), 706–713. https://doi.org/10.1111/ele.12449

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software*, *33*(2), 1–22. https://doi.org/10.18637/jss.v033.i02

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109. https://doi.org/10.1093/biomet/57.1.97

Hodel, R. G. J., Chen, S., Payton, A. C., McDaniel, S. F., Soltis, P., & Soltis, D. E. (2017). Adding loci improves phylogeographic resolution in red mangroves despite increased missing data: Comparing microsatellites and RAD-Seq and investigating loci filtering. *Scientific Reports*, *7*(1), 17598. https://doi.org/10.1038/s41598-017-16810-7

Holeski, L., Monnahan, P., Koseva, B., McCool, N., Lindroth, R. L., & Kelly, J. K. (2014). A high-resolution genetic map of yellow monkeyflower identifies chemical defense QTLs and recombination rate variation. *G3: Genes|Genomes|Genetics*, *4*(5), 813–821. https://doi.org/10.1534/g3.113.010124

Holsinger, K. E. (1988). Inbreeding depression doesn't matter: The genetic basis of mating-system evolution. *Evolution*, *42*, 1235–1244. https://doi.org/10.1111/j.1558-5646.1988.tb04183.x

Jeffries, D. L., Copp, G. H., Lawson Handley, L., Olsén, K. H., Sayer, C. D., & Hänfling, B. (2016). Comparing RADseq and microsatellites to infer complex phylogeographic patterns, an empirical perspective in the Crucian carp, *Carassius carassius*, L. *Molecular Ecology*, *25*(13), 2997–3018. https://doi.org/10.1111/mec.13613

Kalisz, S., & Vogler, D. W. (2003). Benefits of autonomous selfing under unpredictable pollinator environments. *Ecology*, *84*(11), 2928–2942. https://doi.org/10.1890/02-0519

Kelly, J. K. (1999a). Response to selection in partially self fertilizing populations. 2. Selection on multiple traits. *Evolution*, *53*, 350–357.

Kelly, J. K. (1999b). Response to selection in partially self fertilizing populations. I. Selection on a single trait. *Evolution*, *53*, 336–349. https://doi.org/10.1111/j.1558-5646.1999.tb03770.x

Kempthorne, O. (1957). *An introduction to genetic statistics*. New York, NY: Wiley.

Koelling, V. A., Monnahan, P. J., & Kelly, J. K. (2012). A Bayesian method for the joint estimation of outcrossing rate and inbreeding depression. *Heredity*, *109*(6), 393–400. https://doi.org/10.1038/hdy.2012.58

Koboldt, D. C., Chen, K., Wylie, T., Larson, D. E., McLellan, M. D., Mardis, E. R., … Ding, L. (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, *25*(17), 2283–2285. https://doi.org/10.1093/bioinformatics/btp373

Kudo, G., & Ida, T. Y. (2013). Early onset of spring increases the phenological mismatch between plants and pollinators. *Ecology*, *94*(10), 2311–2320. https://doi.org/10.1890/12-2003.1

Ladd, P. G., Thavornkanlapachai, R., & Byrne, M. (2018). Population density and size influence pollen dispersal pattern and mating system of the predominantly outcrossed Banksia nivea (Proteaceae) in a threatened ecological community. *Biological Journal of the Linnean Society*, *124*(3), 492–503. https://doi.org/10.1093/biolinnean/bly050

Lande, R., & Schemske, D. W. (1985). The evolution of self-fertilization and inbreeding depression in plants. I. Genetic models. *Evolution*, *39*(1), 24–40.

Lee, Y. W., Fishman, L., Kelly, J. K., & Willis, J. H. (2016). A segregating inversion generates fitness variation in yellow monkeyflower (*Mimulus guttatus*). *Genetics*, *202*(4), 1473–1484. https://doi.org/10.1534/genetics.115.183566

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arxiv.org/abs/1303.3997v2

Li, H., Handsaker, A., Wysoker, T., Fennell, J., Ruan, N., Homer, G., … G.P.D.P. Subgroup (2009). The sequence alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079.

Li, H., & Ralph, P. (2018). Local PCA shows how the effect of population structure differs along the genome. *bioRxiv*, 070615. https://doi.org/10.1101/070615

Lloyd, D. G. (1979). Some reproductive factors affecting the selection of self-fertilization in plants. *American Naturalist*, *113*, 67–79. https://doi.org/10.1086/283365

Lowry, D. B., & Willis, J. H. (2010). A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLOS Biology*, *8*(9), e1000500. https://doi.org/10.1371/journal.pbio.1000500

Mastretta-Yanes, A., Arrigo, N., Alvarez, N., Jorgensen, T. H., Piñero, D., & Emerson, B. C. (2015). Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources*, *15*(1), 28–41. https://doi.org/10.1111/1755-0998.12291

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092. https://doi.org/10.1063/1.1699114

Mignot, A., Petit, C., Hatt, C., Flaven, É., Pouzadoux, J., Ronce, O., … David, P. (2016). Lower selfing rates in metallicolous populations than in non-metallicolous populations of the pseudometallophyte *Noccaea caerulescens* (Brassicaceae) in Southern France. *Annals of Botany*, *117*(3), 507–519. https://doi.org/10.1093/aob/mcv191

Monnahan, P. J., Colicchio, J., & Kelly, J. K. (2015). A genomic selection component analysis characterizes migration-selection balance. *Evolution*, *69*(7), 1713–1727. https://doi.org/10.1111/evo.12698

Monnahan, P. J., & Kelly, J. K. (2017). The genomic architecture of flowering time varies across space and time in *Mimulus guttatus*. *Genetics*, *206*, 1621–1635. https://doi.org/10.1534/genetics.117.201483

Müller, H. (1883). *The fertilisation of flowers*. London, UK: MacMillan.

Nagamitsu, T., Shuri, K., Taki, H., Kikuchi, S., & Masaki, T. (2016). Effects of converting natural forests to coniferous plantations on fruit and seed production and mating patterns in wild cherry trees. *Ecological Research*, *31*(2), 239–250. https://doi.org/10.1007/s11284-015-1331-x

Nicholson, G., Smith, A., Jonsson, F., Gustafsson, O., Stefanson, K., & Donnelly, P. (2002). Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society: Series B*, *64*, 695–715. https://doi.org/10.1111/1467-9868.00357

Nishizawa, M., & Ohara, M. (2018). The role of sexual and vegetative reproduction in the population maintenance of a monocarpic perennial herb, *Cardiocrinum cordatum* var. glehnii. *Plant Species Biology*, *33*(4), 289–304. https://doi.org/10.1111/1442-1984.12223

Palma-Silva, C., Cozzolino, S., Paggi, G. M., Lexer, C., & Wendt, T. (2015). Mating system variation and assortative mating of sympatric bromeliads (*Pitcairnia* spp.) endemic to neotropical inselbergs. *American Journal of Botany*, *102*(5), 758–764. https://doi.org/10.3732/ajb.1400513

Pannell, J. R., & Voillemot, M. (2017). *Evolution and Ecology of Plant Mating Systems. In* eLS.

Parchman, T. L., Gompert, Z., Mudge, J., Schilkey, F. D., Benkman, C. W., & Buerkle, C. A. (2012). Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecology*, *21*(12), 2991–3005. https://doi.org/10.1111/j.1365-294X.2012.05513.x

Ritland, K. (1990). Inferences about inbreeding depression based on changes of the inbreeding coefficient. *Evolution*, *44*(5), 1230–1241. https://doi.org/10.1111/j.1558-5646.1990.tb05227.x

Ritland, K. (2002). Extensions of models for the estimation of mating systems using n independent loci. *Heredity*, *88*(4), 221–228. https://doi.org/10.1038/sj.hdy.6800029

Ritland, K., & Jain, S. (1981). A model for the estimation of outcrossing rate and gene frequencies using independent loci. *Heredity*, *47*, 35–52. https://doi.org/10.1038/hdy.1981.57

Rivière, E., Lebreton, G., Reynaud, B., Cuénin, N., Flores, O., & Martos, F. (2019). Great genetic diversity but high selfing rates and short-distance gene flow characterize populations of a tree (Foetidia; Lecythidaceae) in the fragmented tropical dry forest of the Mascarene Islands. *Journal of Heredity*, *110*, 287–299. https://doi.org/10.1093/jhered/esy069

Ruegg, K., Anderson, E. C., Boone, J., Pouls, J., & Smith, T. B. (2014). A role for migration-linked genes and genomic islands in divergence of a songbird. *Molecular Ecology*, *23*(19), 4757–4769. https://doi.org/10.1111/mec.12842

Scofield, D., & Schultz, S. (2006). Mitosis, stature and evolution of plant mating systems: Low-Phi and high-Phi plants. *Proceedings of the Royal Society B-Biological Sciences*, *273*, 275–282. https://doi.org/10.1098/rspb.2005.3304

Soria-Carrasco, V., Gompert, Z., Comeault, A. A., Farkas, T. E., Parchman, T. L., Johnston, J. S., … Nosil, P. (2014). Stick insect genomes reveal natural selection's role in parallel speciation. *Science*, *344*(6185), 738–742. https://doi.org/10.1126/science.1252136

Thomson, J. D., & Plowright, R. C. (1980). Pollen carryover, nectar rewards, and pollinator behavior with special reference to *Diervilla lonicera*. *Oecologia*, *46*(1), 68–74. https://doi.org/10.1007/bf00346968

Thomson, J. D., Slatkin, M., & Thomson, B. A. (1997). Trapline foraging by bumble bees. 2. Definition and detection from sequence data. *Behavioral Ecology*, *8*(2), 199–210.

Troth, A., Puzey, J. R., Kim, R. S., Willis, J. H., & Kelly, J. K. (2018). Selective trade-offs maintain alleles underpinning complex trait variation in plants. *Science*, *361*(6401), 475–478. https://doi.org/10.1126/science.aat5760

Twyford, A. D., & Friedman, J. (2015). Adaptive divergence in the monkey flower *Mimulus guttatus* is maintained by a chromosomal inversion. *Evolution*, *69*(6), 1476–1486. https://doi.org/10.1111/evo.12663

Uyenoyama, M. K., Holsinger, K. E., & Waller, D. M. (1993). Ecological and genetic factors directing the evolution of self-fertilization. *Oxford Surveys in Evolutionary Biology*, *9*, 327–381.

Van Etten, M. L., Tate, J. A., Anderson, S. H., Kelly, D., Ladley, J. J., Merrett, M. F., … Robertson, A. W. (2015). The compounding effects of high pollen limitation, selfing rates and inbreeding depression leave a New Zealand tree with few viable offspring. *Annals of Botany*, *116*(5), 833–843. https://doi.org/10.1093/aob/mcv118

Wessinger, C. A., Kelly, J. K., Jiang, P., Rausher, M. D., & Hileman, L. C. (2018). SNP-skimming: A fast approach to map loci generating quantitative variation in natural populations. *Molecular Ecology Resources*, *18*(6), 1402–1414. https://doi.org/10.1111/1755-0998.12930

Whitehead, M. R., Lanfear, R., Mitchell, R. J., & Karron, J. D. (2018). Plant mating systems often vary widely among populations. *Frontiers in Ecology and Evolution*, *6*(38), 1–9. https://doi.org/10.3389/fevo.2018.00038

Willis, J. H. (1993). Effects of different levels of inbreeding on fitness components in Mimulus guttatus. *Evolution*, *47*, 864–876.

Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics*, *15*, 323–354. https://doi.org/10.1111/j.1469-1809.1949.tb02451.x

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, *88*(1), 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011

Yang, M.-L., Wang, L.-L., Zhang, G.-P., Meng, L.-H., Yang, Y.-P., & Duan, Y.-W. (2018). Equipped for migrations across high latitude regions? Reduced spur length and outcrossing rate in a biennial *Halenia elliptica* (Gentianaceae) with mixed mating system along a latitude gradient. *Frontiers in Genetics*, *9*, 223. https://doi.org/10.3389/fgene.2018.00223

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., … Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, *38*(2), 203–208. https://doi.org/10.1038/ng1702

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Colicchio J, Monnahan PJ, Wessinger CA, Brown K, Kern JR, Kelly JK. Individualized mating system estimation using genomic data. *Mol Ecol Resour*. 2020;20:333–347. https://doi.org/10.1111/1755-0998.13094