A Bayesian Approach to Inferring Rates of Selfing and Locus-Specific Mutation

Benjamin D. Redelings,* Seiji Kumagai,* Andrey Tatarenkov,† Liuyang Wang,* Ann K. Sakai,† Stephen G. Weller,† Theresa M. Culley,‡ John C. Avise,† and Marcy K. Uyenoyama*,[1]

*Department of Biology, Duke University, Durham, North Carolina 27708-0338, †Department of Ecology and Evolutionary Biology, University of California, Irvine, California 92697-2525, and ‡Department of Biological Sciences, University of Cincinnati, Cincinnati, Ohio 45220

ORCID IDs: 0000-0002-3278-4343 (B.D.R.); 0000-0002-0516-5862 (A.T.); 0000-0001-9556-2361 (L.W.); 0000-0001-8249-1103 (M.K.U.)

ABSTRACT We present a Bayesian method for characterizing the mating system of populations reproducing through a mixture of self-fertilization and random outcrossing. Our method uses patterns of genetic variation across the genome as a basis for inference about reproduction under pure hermaphroditism, gynodioecy, and a model developed to describe the self-fertilizing killifish *Kryptolebias marmoratus*. We extend the standard coalescence model to accommodate these mating systems, accounting explicitly for multilocus identity disequilibrium, inbreeding depression, and variation in fertility among mating types. We incorporate the Ewens sampling formula (ESF) under the infinite-alleles model of mutation to obtain a novel expression for the likelihood of mating system parameters. Our Markov chain Monte Carlo (MCMC) algorithm assigns locus-specific mutation rates, drawn from a common mutation rate distribution that is itself estimated from the data using a Dirichlet process prior model. Our sampler is designed to accommodate additional information, including observations pertaining to the sex ratio, the intensity of inbreeding depression, and other aspects of reproduction. It can provide joint posterior distributions for the population-wide proportion of uniparental individuals, locus-specific mutation rates, and the number of generations since the most recent outcrossing event for each sampled individual. Further, estimation of all basic parameters of a given model permits estimation of functions of those parameters, including the proportion of the gene pool contributed by each sex and relative effective numbers.

KEYWORDS selfing rate; Ewens sampling formula; Bayesian; MCMC; mating system

INBREEDING generates genome-wide, multilocus disequilibria of various orders, transforming the context in which evolution proceeds. Here, we address a simple form of inbreeding: a mixture of self-fertilization (selfing) and random outcrossing (Clegg 1980; Ritland 2002).

Various methods exist for the estimation of selfing rates from genetic data. Wright's (1921) fundamental approach bases the estimation of selfing rates on the coefficient of inbreeding (F_{IS}), a summary of the departure from Hardy–Weinberg proportions of genotypes for a given set of allele frequencies. The maximum-likelihood method of Enjalbert and David (2000) detects inbreeding from departures of multiple unlinked loci from Hardy–Weinberg proportions, estimating allele frequencies for each locus and accounting for correlations in heterozygosity among loci [identity disequilibrium (Cockerham and Weir 1968)]. David *et al.* (2007) extend the approach of Enjalbert and David (2000) to accommodate errors in scoring heterozygotes as homozygotes. A primary objective of InStruct (Gao *et al.* 2007) is the estimation of admixture. It extends the widely used program structure (Pritchard *et al.* 2000), which bases the estimation of admixture on disequilibria of various forms, by accounting for disequilibria due to selfing. Progeny array methods (see Ritland 2002), which base the estimation of selfing rates on the genetic analysis of family data, are particularly well suited to plant populations. Wang *et al.* (2012) extend this approach to a random sample of individuals by reconstructing sibship relationships within the sample.

Methods that base the estimation of inbreeding rates on the observed departure from random union of gametes require

doi: 10.1534/genetics.115.179093

Manuscript received June 5, 2015; accepted for publication September 4, 2015; published Early Online September 14, 2015.

Supporting information is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.179093/-/DC1.

[1]Corrresponding author: Department of Biology, Box 90338, Duke University, Durham, NC 27708-0338. E-mail: marcy@duke.edu

information on expected Hardy–Weinberg proportions. Population-wide frequencies of alleles observed in a sample at locus l ($\{p_{li}\}$) can be estimated jointly in a maximum-likelihood framework (*e.g.*, Hill *et al.* 1995) or integrated out as nuisance parameters in a Bayesian framework (*e.g.*, Ayres and Balding 1998). Similarly, expected locus-specific heterozygosity,

$$d_l = 1 - \sum_i p_{li}^2, \tag{1}$$

can be obtained from observed allele frequencies (Enjalbert and David 2000) or estimated jointly with the selfing rate (David *et al.* 2007).

Here, we introduce a Bayesian method for the analysis of mixed-mating systems that accounts for genetic variation through coalescence-based models and uses the Ewens sampling formula (ESF) (Ewens 1972) in determining likelihoods. Our approach replaces the estimation of allele frequencies or heterozygosity (Equation 1) with the estimation of a locus-specific mutation rate (θ^*) under the infinite-alleles model of mutation. We use a Dirichlet process prior (DPP) to determine the number of classes of mutation rates, the mutation rate for each class, and the class membership of each locus. We assign the DPP parameters in a conservative manner so that a new mutational class is created only if sufficient evidence exists to justify doing so. Further, while other methods assume that the frequency in the population of an allelic class not observed in the sample is zero, the ESF provides the probability, under the infinite-alleles model of mutation, that the next-sampled gene represents a novel allele [see (21a)].

To estimate the probability that a random individual is uniparental (s^*), we exploit identity disequilibrium (Cockerham and Weir 1968), the correlation in heterozygosity across loci. This association, even among unlinked loci, reflects that all loci within an individual share a history of inbreeding back to the most recent random outcrossing event. Conditional on the number of generations since this event, the genealogical histories of unlinked loci are independent. For each diploid individual in the sample, our method models coalescence events at each locus back to the most recent point at which all remaining lineages reside in distinct individuals. The ESF provides the exact likelihood of the ancestral allele frequency spectrum at that point, obviating the need for further genealogical reconstruction. This approach permits computationally efficient analysis of samples comprising large numbers of individuals and large numbers of loci observed across the genome.

We address the estimation of rates of inbreeding and other evolutionary processes in populations undergoing pure hermaphroditism, androdioecy (hermaphrodites and males), or gynodioecy (hermaphrodites and females). Application of the method to simulated data sets demonstrates its accuracy in parameter estimation and in assessing uncertainty. We apply the method to microsatellite data from the self-fertilizing killifish *Kryptolebias marmoratus* (Mackiewicz *et al.* 2006; Tatarenkov *et al.* 2012) and the gynodioecious Hawaiian endemic *Schiedea salicaria* (Wallace *et al.* 2011) to illustrate the simultaneous inference of various biologically significant aspects of mating systems in nature, including levels of inbreeding depression, population proportions of sexual forms, and effective numbers.

Evolutionary Model

We use the ESF (Ewens 1972) to determine likelihoods based on a sample of diploid multilocus genotypes. By subsampling a single gene from each locus from each diploid individual, we could apply the ESF to the reduced sample to determine a likelihood function with a single parameter: the mutation rate, appropriately scaled to account for the acceleration of the coalescence rate caused by inbreeding [θ^* (Fu 1997; Nordborg and Donnelly 1997)]. Consideration of the full sample of diploid genotypes yields information about an additional parameter: the probability that a random individual is uniparental (uniparental proportion s^*).

We describe the dependence of composite parameters s^* and θ^* on the basic parameters of the iconic mating systems pure hermaphroditism and gynodioecy. In addition, we develop the *Kryptolebias* model, based on the mating system of the killifish *K. marmoratus*, in which only males fertilize eggs that are not self-fertilized by hermaphrodites (Furness *et al.* 2015). Although this mating system and that of the worm *Caenorhabditis elegans* have been described as androdioecious, we reserve this botanical term for plant systems comprising hermaphrodites and female steriles (males), with pollen from both sexes capable of fertilizing seeds that are not set by self-pollen.

Rates of coalescence and mutation

Here, we describe the structure of the coalescence process shared by our pure hermaphroditism, *Kryptolebias*, and gynodioecy models.

Relative rates of coalescence and mutation: We use s^* to denote the uniparental proportion (probability that a random individual is uniparental) and $1/N^*$ to denote the rate of parent sharing (the probability that a pair of genes residing in distinct individuals descend from the same individual in the immediately preceding generation). These quantities determine the coalescence rate and the scaled mutation rate of the ESF.

A pair of lineages residing in distinct individuals derive from a single parent (P) in the preceding generation at rate $1/N^*$. They descend from the same gene (immediate coalescence) or from distinct genes in P with equal probability. In the latter case, P is itself either uniparental (probability s^*), implying descent once again of the lineages from a single individual in the preceding generation, or biparental, implying descent from distinct individuals. The ancestry of a pair of lineages residing in a single individual rapidly resolves either to coalescence, with probability

$$f_c = \frac{s^*}{2 - s}$$

[the classical coefficient of identity (Wright 1921; Haldane 1924)], or to residence in distinct individuals, with the complement probability. The total rate of coalescence of lineages sampled from distinct individuals corresponds to

$$\frac{(1 + f_c)/2}{N^*} = \frac{1}{N^*(2 - s^*)}. \tag{2}$$

Our model assumes that coalescence and mutation occur on comparable timescales,

$$\lim_{\substack{N \to \infty \\ u \to 0}} 4Nu = \theta$$

$$\lim_{\substack{N \to \infty \\ N^* \to \infty}} \frac{N^*}{N} = E, \tag{3}$$

for u the rate of mutation under the infinite-alleles model and N an arbitrary quantity that goes to infinity at a rate comparable to N^* and $1/u$. Here, E represents a measure of effective population size (the "inbreeding effective size" of Crow and Denniston 1988), scaled relative to a population comprising N reproductives.

In large populations, switching of lineages between uniparental and biparental carriers occurs on the order of generations, virtually instantaneously relative to the rate at which lineages residing in distinct individuals coalesce (Fu 1997; Nordborg and Donnelly 1997). Our model assumes independence between the processes of coalescence and mutation and that these processes occur on a much longer timescale than random outcrossing:

$$1 - s^* \gg u, \frac{1}{N^*}. \tag{4}$$

Using (2), we obtain the probability that the most recent event in the ancestry of m lineages, each residing in a distinct individual, corresponds to mutation,

$$\lim_{N \to \infty} \frac{um}{um + \binom{m}{2} \Big/ \left[N^*(2 - s^*)\right]} = \frac{\theta^*}{\theta^* + m - 1},$$

in which

$$\theta^* = \lim_{\substack{N \to \infty \\ u \to 0}} 2N^* u(2 - s^*) = \lim_{\substack{N \to \infty \\ u \to 0}} 4Nu \frac{N^*}{N} \left(1 - s^* \big/ 2\right)$$

$$= \theta(1 - s^*/2)E, \tag{5}$$

for θ and E defined in (3). In inbred populations, the single parameter of the ESF for an allele frequency spectrum comprising genes sampled from separate individuals corresponds to θ^*.

Uniparental proportion and the rate of parent sharing: In a purely hermaphroditic population comprising N_h reproductives, the rate of parent sharing ($1/N^*$) corresponds to $1/N_h$ and the uniparental proportion (s^*) to

$$s_H = \frac{\tilde{s}\tau}{\tilde{s}\tau + 1 - \tilde{s}}, \tag{6a}$$

for $\tilde{s}$ the fraction of uniparental offspring at conception and τ the rate of survival of uniparental relative to biparental offspring. For the pure-hermaphroditism model, we assign the arbitrary constant N in (3) as N_h, implying

$$E_H = \frac{N_h}{N} \equiv 1. \tag{6b}$$

Under the *Kryptolebias* model, involving reproduction by N_h hermaphrodites and N_m males, the uniparental proportion (s^*) is identical to the case of pure hermaphroditism (Equation 6),

$$s_L = \frac{\tilde{s}\tau}{\tilde{s}\tau + 1 - \tilde{s}}. \tag{7a}$$

Because only males fertilize eggs that are not self-fertilized by hermaphrodites, a random gene derives from a male in the preceding generation with probability

$$\frac{1 - s_L}{2}.$$

The rate of parent sharing ($1/N^*$) corresponds to

$$\frac{1}{N_L} = \frac{\left[(1 + s_L)/2\right]^2}{N_h} + \frac{\left[(1 - s_L)/2\right]^2}{N_m}, \tag{7b}$$

which in the absence of inbreeding ($s_L = 0$) agrees with the classical harmonic mean expression for effective population size (Wright 1969). For the *Kryptolebias* model, we assign the arbitrary constant N in (3) as the number of reproductives ($N_h + N_m$), implying a scaled rate of coalescence of

$$\frac{1}{E_L} = \frac{N_h + N_m}{N_L} = \frac{\left[(1 + s_L)/2\right]^2}{1 - p_m} + \frac{\left[(1 - s_L)/2\right]^2}{p_m}, \tag{7c}$$

for

$$p_m = \frac{N_m}{N_h + N_m}, \tag{8a}$$

the proportion of males among reproductives. Relative effective number $E_L \in (0, 1]$ takes its maximum under equality between the total number of reproductives ($N_h + N_m$) and effective number N_L, determined by the rate of parent sharing. At $E_L = 1$, the probability that a random gene derives from a male parent corresponds to the proportion of males among reproductives:

$$\frac{1-s_{\mathrm{L}}}{2}=p_{\mathrm{m}}. \tag{8b}$$

In gynodioecious populations, in which N_{h} hermaphrodites and N_{f} females (male steriles) reproduce, the uniparental proportion (s^*) corresponds to

$$s_{\mathrm{G}}=\frac{\tau N_{\mathrm{h}}\tilde{s}}{\tau N_{\mathrm{h}}\tilde{s}+N_{\mathrm{h}}(1-\tilde{s})+N_{\mathrm{f}}\sigma}, \tag{9a}$$

in which σ represents the seed fertility of females relative to hermaphrodites and $\tilde{s}$ is the proportion of seeds of hermaphrodites set by self-pollen. A random gene derives from a female in the preceding generation with probability

$$\frac{(1-s_{\mathrm{G}})F}{2},$$

for

$$F=\frac{N_{\mathrm{f}}\sigma}{N_{\mathrm{h}}(1-\tilde{s})+N_{\mathrm{f}}\sigma}, \tag{9b}$$

the proportion of biparental offspring that have a female parent. The rate of parent sharing ($1/N^*$) corresponds to

$$\frac{1}{N_{\mathrm{G}}}=\frac{\left[1-(1-s_{\mathrm{G}})F/2\right]^2}{N_{\mathrm{h}}}+\frac{\left[(1-s_{\mathrm{G}})F/2\right]^2}{N_{\mathrm{f}}}. \tag{9c}$$

We assign the arbitrary constant N in (3) as $(N_{\mathrm{h}}+N_{\mathrm{f}})$, implying a scaled rate of coalescence of

$$\frac{1}{E_{\mathrm{G}}}=\frac{N_{\mathrm{h}}+N_{\mathrm{f}}}{N_{\mathrm{G}}}=\frac{\left[1-(1-s_{\mathrm{G}})F/2\right]^2}{1-p_{\mathrm{f}}}+\frac{\left[(1-s_{\mathrm{G}})F/2\right]^2}{p_{\mathrm{f}}}, \tag{9d}$$

for

$$p_{\mathrm{f}}=\frac{N_{\mathrm{f}}}{N_{\mathrm{h}}+N_{\mathrm{f}}}, \tag{10a}$$

the proportion of females among reproductives. As for the *Kryptolebias* model, $E_{\mathrm{G}}\in(0,1]$ achieves its maximum only if the proportion of females among reproductives equals the probability that a random gene derives from a female parent:

$$\frac{(1-s_{\mathrm{G}})F}{2}=p_{\mathrm{f}}. \tag{10b}$$

Likelihood

We here address the probability of a sample of diploid multilocus genotypes.

Genealogical histories: For a sample comprising up to two alleles at each of L autosomal loci in n diploid individuals, we represent the observed genotypes by

$$\mathbf{X}=\{\mathbf{X}_1,\mathbf{X}_2,\ldots,\mathbf{X}_L\}, \tag{11}$$

in which $\mathbf{X}_l$ denotes the set of genotypes observed at locus l among the n individuals

$$\mathbf{X}_l=\{\mathbf{X}_{l1,}\mathbf{X}_{l2},\ldots,\mathbf{X}_{ln}\}, \tag{12}$$

with

$$\mathbf{X}_{lk}=\left(X_{lk1},X_{lk2}\right)$$

the genotype at locus l of individual k, which bears alleles X_{lk1} and X_{lk2}.

To facilitate accounting for the shared recent history of genes borne by an individual in sample, we introduce latent variables

$$\mathbf{T}=\{T_1,T_2,\ldots,T_n\}, \tag{13}$$

for T_k denoting the number of consecutive generations of selfing in the immediate ancestry of the kth individual, and

$$\mathbf{I}=\{I_{lk}\}, \tag{14}$$

for I_{lk} indicating whether the lineages borne by the kth individual at locus l coalesce within the most recent T_k generations. Independent of other individuals, the number of consecutive generations of inbreeding in the ancestry of the kth individual is geometrically distributed,

$$T_k\sim\text{Geometric}\left(s^*\right), \tag{15}$$

with $T_k=0$ signifying that individual k is the product of random outcrossing. Irrespective of whether 0, 1, or 2 of the genes at locus l in individual k are observed, I_{lk} indicates whether the two genes at that locus in individual k coalesce during the T_k consecutive generations of inbreeding in its immediate ancestry:

$$I_{lk}=\left\{\begin{array}{ll}0 & \text{if the two genes do not coalesce}\\ 1 & \text{if the two genes coalesce.}\end{array}\right.$$

Because the pair of lineages at any locus coalesce with probability 1/2 in each generation of selfing,

$$\Pr\left(I_{lk}=0\right)=\frac{1}{2^{T_k}}=1-\Pr\left(I_{lk}=1\right). \tag{16}$$

Figure 1 depicts the recent genealogical history at a locus l in five individuals. Individuals 2 and 5 are products of random outcrossing ($T_2=T_5=0$), while the others derive from some positive number of consecutive generations of selfing in their immediate ancestry ($T_1=2,T_3=3,T_4=1$). Both individuals 1 and 3 are homozygotes ($\alpha\alpha$), with the lineages of individual 3 but not 1 coalescing more recently than the most recent outcrossing event ($I_{l1}=0,I_{l3}=1$). As individual 2 is heterozygous ($\alpha\beta$), its lineages necessarily remain distinct since the most recent outcrossing event ($I_{l2}=0$). One gene in each of individuals 4 and 5 is unobserved (*), with the unobserved

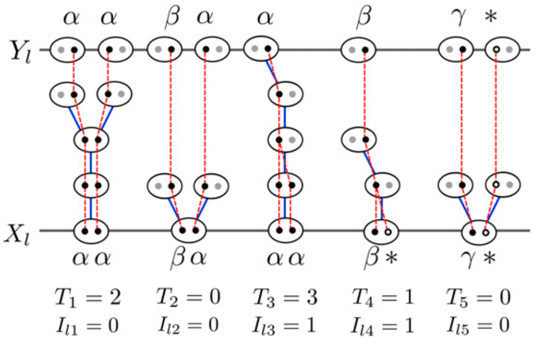

α α β α α β γ *
Y_l
X_l
αα βα αα β∗ γ∗
$T_1 = 2$ $T_2 = 0$ $T_3 = 3$ $T_4 = 1$ $T_5 = 0$
$I_{l1} = 0$ $I_{l2} = 0$ $I_{l3} = 1$ $I_{l4} = 1$ $I_{l5} = 0$

Figure 1 Following the history of the sample ($\mathbf{X}_l$) backward in time until all ancestors of sampled genes reside in different individuals ($\mathbf{Y}_l$). Ovals represent individuals and circles represent genes. Blue lines indicate the parents of individuals, while red lines represent the ancestry of genes. Black circles represent sampled genes for which the allelic class is observed (Greek letters) and their ancestral lineages. White circles represent genes in the sample with unobserved allelic class (*). Gray circles represent other genes carried by ancestors of the sampled individuals. The relationship between the observed sample $\mathbf{X}_l$ and the ancestral sample $\mathbf{Y}_l$ is determined by the intervening coalescence events $\mathbf{I}_l$. $\mathbf{T}$ indicates the number of consecutive generations of selfing for each sampled individual.

lineage in individual 4 but not 5 coalescing more recently than the most recent outcrossing event ($I_{l4} = 1, I_{l5} = 0$).

In addition to the observed sample of diploid individuals, we consider the state of the sampled lineages at the most recent generation in which an outcrossing event has occurred in the ancestry of all n individuals. This point in the history of the sample occurs $\hat{T}$ generations into the past, for

$$\hat{T} = 1 + \max_k T_k.$$

In Figure 1, for example, $\hat{T} = 4$, reflecting the most recent outcrossing event in the ancestry of individual 3. As all remaining lineages reside in distinct individuals at that point, the ESF provides the probability of the allele frequency spectrum at this point.

We represent the ordered list of allelic states of the lineages at $\hat{T}$ generations into the past by

$$\mathbf{Y} = \{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_L\}, \tag{17}$$

for $\mathbf{Y}_l$ a list of ancestral genes in the same order as their descendants in $\mathbf{X}_l$. Each gene in $\mathbf{Y}_l$ is the ancestor of either 1 or 2 genes at locus l from a particular individual in $\mathbf{X}_l$ (Equation 12), depending on whether the lineages held by that individual coalesce during the consecutive generations of inbreeding in its immediate ancestry. We represent the number of genes in $\mathbf{Y}_l$ by m_l ($n \le m_l \le 2n$). In Figure 1, for example, $\mathbf{X}_l$ contains 10 genes in five individuals, but $\mathbf{Y}_l$ contains only 8 genes, with Y_{l1} the ancestor of only the first allele of $\mathbf{X}_{l1}$ and Y_{l5} the ancestor of both alleles of $\mathbf{X}_{l3}$.

We assume (Equation 4) that the initial phase of consecutive generations of selfing is sufficiently short to ensure a negligible probability of mutation in any lineage at any locus and a negligible probability of coalescence between lineages held by distinct individuals more recently than $\hat{T}$. In addition to constraints on relative rates within loci (Equation 4), this assumption may entail small numbers of observed loci relative to the population size ($n \ll N^*$). Under these assumptions, the coalescence history $\mathbf{I}$ (Equation 14) completely determines the correspondence between genetic lineages in $\mathbf{X}$ (Equation 11) and $\mathbf{Y}$ (Equation 17).

Computing the likelihood: In principle, the likelihood of the observed data can be computed from the augmented likelihood by summation,

$$\Pr\left(\mathbf{X}|\boldsymbol{\Theta}^*, s^*\right) = \sum_{\mathbf{I}} \sum_{\mathbf{T}} \Pr\left(\mathbf{X}, \mathbf{I}, \mathbf{T}|\boldsymbol{\Theta}^*, s^*\right), \tag{18}$$

for

$$\boldsymbol{\Theta}^* = \left\{\theta_1^*, \theta_2^*, \ldots, \theta_L^*\right\}, \tag{19}$$

the list of scaled, locus-specific mutation rates, s^* the population-wide uniparental proportion for the reproductive system under consideration (*e.g.*, Equation 6 for the pure hermaphroditism model), and $\mathbf{T}$ (Equation 13) and $\mathbf{I}$ (Equation 14) the lists of latent variables representing the time since the most recent outcrossing event and whether the two lineages borne by a sampled individual coalesce during this period. Here we follow a common abuse of notation in using $\Pr(\mathbf{X})$ to denote $\Pr(\mathbf{X} = \mathbf{x})$ for random variable $\mathbf{X}$ and realized value $\mathbf{x}$. Summation (18) is computationally expensive: the number of consecutive generations of inbreeding in the immediate ancestry of an individual (T_k) has no upper limit (compare David *et al.* 2007) and the number of combinations of coalescence states (I_{lk}) across the L loci and n individuals increases exponentially (2^{Ln}) with the total number of assignments. We perform Markov chain Monte Carlo (MCMC) to avoid both these sums.

To calculate the augmented likelihood, we begin by applying Bayes' rule:

$$\Pr\left(\mathbf{X}, \mathbf{I}, \mathbf{T}|\boldsymbol{\Theta}^*, s^*\right) = \Pr\left(\mathbf{X}, \mathbf{I}|\mathbf{T}, \boldsymbol{\Theta}^*, s^*\right)\Pr\left(\mathbf{T}|\boldsymbol{\Theta}^*, s^*\right).$$

Because the times since the most recent outcrossing event $\mathbf{T}$ depend only on the uniparental proportion s^*, through (15), and not on the rates of mutation $\boldsymbol{\Theta}^*$,

$$\Pr\left(\mathbf{T}|\boldsymbol{\Theta}^*, s^*\right) = \prod_{k=1}^{n} \Pr\left(T_k|s^*\right).$$

Even though our model assumes the absence of physical linkage among any of the loci, the genetic data $\mathbf{X}$ and coalescence events $\mathbf{I}$ are not independent across loci because they depend on the times since the most recent outcrossing event $\mathbf{T}$. Given $\mathbf{T}$, however, the genetic data and coalescence events are independent across loci:

$$\Pr\left(\mathbf{X}, \mathbf{I}|\mathbf{T}, \boldsymbol{\Theta}^*, s^*\right) = \prod_{l=1}^{L} \Pr\left(\mathbf{X}_l, \mathbf{I}_l|\mathbf{T}, \theta_l^*, s^*\right).$$

Further,

$$\begin{aligned}\Pr\big(\mathbf{X}_l,\mathbf{I}_l|\mathbf{T},\theta_l^*,s^*\big) &= \Pr\big(\mathbf{X}_l|\mathbf{I}_l,\mathbf{T},\theta_l^*,s^*\big)\cdot\Pr\big(\mathbf{I}_l|\mathbf{T},\theta_l^*,s^*\big)\\ &= \Pr\big(\mathbf{X}_l|\mathbf{I}_l,\theta_l^*,s^*\big)\cdot\prod_{k=1}^{n}\Pr(I_{lk}|T_k).\end{aligned}$$

This expression reflects that the times to the most recent outcrossing event $\mathbf{T}$ affect the observed genotypes $\mathbf{X}_l$ only through the coalescence states $\mathbf{I}_l$ and that the coalescence states $\mathbf{I}_l$ depend only on the times to the most recent outcrossing event $\mathbf{T}$, through (16).

To compute $\Pr(\mathbf{X}_l|\mathbf{I}_l,\theta_l^*,s^*)$, we incorporate latent variable $\mathbf{Y}_l$ (Equation 17), describing the states of lineages at the most recent point at which all occur in distinct individuals (Figure 1),

$$\begin{aligned}\Pr\Big(\mathbf{X}_l|\mathbf{I}_l,\theta_l^*,s^*\Big) &= \sum_{\mathbf{Y}_l}\Pr\big(\mathbf{X}_l,\ \mathbf{Y}_l|\mathbf{I}_l,\theta_l^*,s^*\big)\\ &= \sum_{\mathbf{Y}_l}\Pr\big(\mathbf{X}_l|\mathbf{Y}_l,\mathbf{I}_l,\theta_l^*,s^*\big)\Pr\big(\mathbf{Y}_l|\mathbf{I}_l,\theta_l^*,s^*\big)\\ &= \sum_{\mathbf{Y}_l}\Pr\big(\mathbf{X}_l|\mathbf{Y}_l,\mathbf{I}_l\big)\cdot\Pr\big(\mathbf{Y}_l|\mathbf{I}_l,\theta_l^*\big),\end{aligned}\tag{20a}$$

reflecting that the coalescence states $\mathbf{I}_l$ establish the correspondence between the spectrum of genotypes in $\mathbf{X}_l$ and the spectrum of alleles in $\mathbf{Y}_l$ and that the distribution of $\mathbf{Y}_l$, given by the ESF, depends on the uniparental proportion s^* only through the scaled mutation rate θ_l^* (Equation 5).

Given the sampled genotypes $\mathbf{X}_l$ and coalescence states $\mathbf{I}_l$, at most one ordered list of alleles $\mathbf{Y}_l$ produces positive $\Pr(\mathbf{X}_l|\mathbf{Y}_l,\ \mathbf{I}_l)$ in (20a). Coalescence of the lineages at locus l in any heterozygous individual [*e.g.*, $X_{lk}=(\beta,\alpha)$ with $I_{lk}=1$ in Figure 1] implies

$$\Pr\big(\mathbf{X}_l|\mathbf{Y}_l,\mathbf{I}_l\big)=0$$

for all $\mathbf{Y}_l$. Any nonzero $\Pr(\mathbf{X}_l|\mathbf{Y}_l,\ \mathbf{I}_l)$ precludes coalescence in any heterozygous individual and $\mathbf{Y}_l$ must specify the observed alleles of $\mathbf{X}_l$ in the order of observation, with either 1 ($I_{lk}=1$) or 2 ($I_{lk}=0$) instances of the allele for any homozygous individual [*e.g.*, $X_{lk}=(\alpha,\alpha)$]. For all cases with nonzero $\Pr(\mathbf{X}_l|\mathbf{Y}_l,\ \mathbf{I}_l)$,

$$\Pr\big(\mathbf{X}_l|\mathbf{Y}_l,\mathbf{I}_l\big)=1.$$

Accordingly, expression (20a) reduces to

$$\Pr\big(\mathbf{X}_l|\mathbf{I}_l,\theta_l^*,s^*\big)=\sum_{\mathbf{Y}_l:\Pr(\mathbf{X}_l|\mathbf{Y}_l,\mathbf{I}_l)\neq 0}\Pr\big(\mathbf{Y}_l|\mathbf{I}_l,\theta_l^*\big),\tag{20b}$$

a sum with either 0 or 1 terms. Because all genes in $\mathbf{Y}_l$ reside in distinct individuals, we obtain $\Pr(\mathbf{Y}_l|\mathbf{I}_l,\theta_l^*)$ from the Ewens sampling formula for a sample, of size

$$m_l=2n-\sum_{k=1}^{n}I_{lk},$$

ordered in the sequence in which the genes are observed.

To determine $\Pr(\mathbf{Y}_l|\mathbf{I}_l,\theta_l^*)$ in (20b), we use a fundamental property of the ESF (Ewens 1972; Karlin and McGregor 1972): the probability that the next-sampled (ith) gene represents a novel allele corresponds to

$$\pi_i=\frac{\theta^*}{i-1+\theta^*},\tag{21a}$$

for θ^* defined in (5), and the probability that it represents an additional copy of already-observed allele j is

$$(1-\pi_i)\frac{i_j}{i-1},\tag{21b}$$

for i_j the number of replicates of allele j in the sample at size $(i-1)$ $(\sum_j i_j=i-1)$. *Appendix A* presents a first-principles derivation of (21a). Expressions (21) imply that for $\mathbf{Y}_l$ the list of alleles at locus l in order of observance,

$$\Pr\big(\mathbf{Y}_l|\mathbf{I}_l,\theta_l^*\big)=\frac{\big(\theta_l^*\big)^{K_l}\prod_{j=1}^{K_l}(m_{lj}-1)!}{\prod_{i=1}^{m_l}\big(i-1+\theta_l^*\big)},\tag{22}$$

in which K_l denotes the total number of distinct allelic classes, m_{lj} the number of replicates of the jth allele in the sample, and $m_l=\sum_j m_{lj}$ the number of lineages remaining at time $\widehat{T}$ (Figure 1).

Missing data: Our method allows the allelic class of one or both genes at each locus to be missing. In Figure 1, for example, the genotype of individual 4 is $\mathbf{X}_{l4}=(\beta,*)$, indicating that the allelic class of the first gene is observed to be β, but that of the second gene is unknown.

A missing allelic specification in the sample of genotypes $\mathbf{X}_l$ leads to a missing specification for the corresponding gene in $\mathbf{Y}_l$ unless the genetic lineage coalesces, in the interval between $\mathbf{X}_l$ and $\mathbf{Y}_l$, with a lineage ancestral to a gene for which the allelic type was observed. Figure 1 illustrates such a coalescence event in the case of individual 4. In contrast, the lineages ancestral to the genes carried by individual 5 fail to coalesce more recently than their separation into distinct individuals, giving rise to a missing specification in $\mathbf{Y}_l$.

The probability of $\mathbf{Y}_l$ can be computed by simply summing over all possible values for each missing specification. Equivalently, those elements may simply be dropped from $\mathbf{Y}_l$ before computing the probability via the ESF, the procedure implemented in our method.

Bayesian Inference Framework

Prior on mutation rates

Ewens (1972) showed for the panmictic case that the number of distinct allelic classes observed at a locus [*e.g.*, K_l in (22)] provides a sufficient statistic for the estimation of the scaled mutation rate. As each locus l provides relatively little information about the scaled mutation rate θ_l^* (Equation 5), we make the assumption that mutation rates across loci cluster in a finite number of groups. Because we do not know *a priori* the group assignment of loci or even the number of distinct rate classes among the observed loci, we use the DPP to estimate simultaneously the number of groups, the value of θ^* for each group, and the assignment of loci to groups.

The Dirichlet process comprises a base distribution, which here represents the distribution of the scaled mutation rate θ^* across groups, and a concentration parameter α, which controls the probability that each successive locus belongs to a new group. In assigning 0.1 to α, which implies a low expected number of rate classes, we adopt a conservative approach under which a new rate class is created only if the data provide sufficient support for doing so. Further, we place a gamma distribution $[\Gamma(\alpha = 0.25, \beta = 2)]$ on the mean scaled mutation rate for each group. As this prior has a high variance relative to the mean (0.5), it is relatively uninformative about θ^*.

Model-specific parameters

Derivations presented in the preceding section indicate that the probability of a sample of diploid genotypes under the infinite-alleles model depends on only the uniparental proportion s^* and the scaled mutation rates $\mathbf{\Theta}^*$ (Equation 19) across loci. These composite parameters are determined by the set of basic demographic parameters $\mathbf{\Psi}$ associated with each model of reproduction under consideration. As the genotypic data provide equal support to any combination of basic parameters that implies the same values of s^* and $\mathbf{\Theta}^*$, the full set of basic parameters for any model is in general nonidentifiable, using the observed genotype frequency spectrum alone.

Even so, our MCMC implementation updates the basic parameters directly, with likelihoods determined from the implied values of s^* and $\mathbf{\Theta}^*$. This feature facilitates the incorporation of information in addition to the genotypic data that can contribute to the estimation of the basic parameters under a particular model or assessment of alternative models. We have

$$\begin{aligned}\Pr(\mathbf{X}, \mathbf{\Theta}^*, \mathbf{\Psi}) &= \Pr(\mathbf{X}|\mathbf{\Theta}^*, \mathbf{\Psi})\cdot\Pr(\mathbf{\Theta}^*)\cdot\Pr(\mathbf{\Psi}) \\ &= \Pr(\mathbf{X}|\mathbf{\Theta}^*, s^*(\mathbf{\Psi}))\cdot\Pr(\mathbf{\Theta}^*)\cdot\Pr(\mathbf{\Psi}),\end{aligned} \tag{23}$$

for $\mathbf{X}$ the genotypic data and $s^*(\mathbf{\Psi})$ the uniparental proportion determined by $\mathbf{\Psi}$ for the model under consideration. To determine the marginal distribution of θ_l (Equation 3) for each locus l, we use (5), incorporating the distributions of $s^*(\mathbf{\Psi})$ and $E(\mathbf{\Psi})$, the scaling factor defined in (3):

$$\theta_l = \frac{\theta_l^*}{E(1 - s^*/2)}.$$

For the pure hermaphroditism model (Equation 6), $\mathbf{\Psi} = \{\tilde{s}, \tau\}$, for $\tilde{s}$ the proportion of conceptions through selfing and τ the viability of uniparental individuals relative to biparental individuals. The default priors for $\tilde{s}$ and τ are uniform:

$$\begin{aligned}\tilde{s} &\sim \text{Uniform}(0, 1) \\ \tau &\sim \text{Uniform}(0, 1).\end{aligned} \tag{24}$$

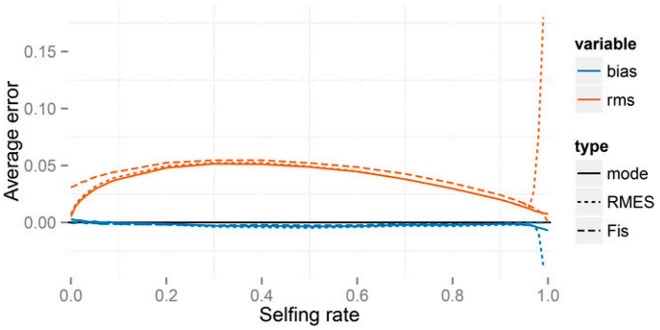


Figure 2 Errors for the full likelihood (posterior mode), RMES, and F_{IS}-based (Equation 27) methods for a large simulated sample ($n = 70$ individuals, $L = 32$ loci). In the key, rms indicates the root-mean-square error and bias the average deviation. Averages are taken across simulated data sets at each true value of s^*.

For the *Kryptolebias* model (Equation 7), $\mathbf{\Psi} = \{\tilde{s}, \tau, p_{\text{m}}\}$, with uniform priors as the default:

$$\begin{aligned}\tilde{s} &\sim \text{Uniform}(0, 1) \\ \tau &\sim \text{Uniform}(0, 1) \\ p_{\text{m}} &\sim \text{Uniform}(0, 1).\end{aligned} \tag{25}$$

For the gynodioecy model (Equation 9), $\mathbf{\Psi} = \{\tilde{s}, \tau, p_{\text{f}}, \sigma\}$, including $\tilde{s}$ the proportion of egg cells produced by hermaphrodites fertilized by selfing, p_{f} (Equation 10a) the proportion of females (male steriles) among reproductives, and σ the fertility of females relative to hermaphrodites. The default priors correspond to

$$\begin{aligned}\tilde{s} &\sim \text{Uniform}(0, 1) \\ \tau &\sim \text{Uniform}(0, 1) \\ p_{\text{f}} &\sim \text{Uniform}(0, 1) \\ \frac{1}{\sigma} &\sim \text{Uniform}(0, 1).\end{aligned} \tag{26}$$

Assessment of Accuracy and Coverage Using Simulated Data

We developed a forward-in-time simulator (https://github.com/skumagai/selfingsim) that tracks multiple neutral loci with locus-specific scaled mutation rates ($\mathbf{\Theta}$) in a population comprising $N = 10^4$ reproducing diploid hermaphrodites of which a proportion s^* are of uniparental origin. We used this simulator to generate data under two sampling regimes: large ($L = 32$ loci in each of $n = 70$ diploid individuals) and small ($L = 6$ loci in each of $n = 10$ diploid individuals). We applied our Bayesian method and RMES (David *et al.* 2007) to simulated data sets. A description of the procedures used to assess the accuracy and coverage properties of the three methods is included in Supporting Information, File S1.

In addition, we determine the uniparental proportion (s^*) inferred from the departure from Hardy–Weinberg expectation (F_{IS}) (Wright 1969) alone. Our F_{IS}-based estimate entails setting the observed value of F_{IS} equal to its classical

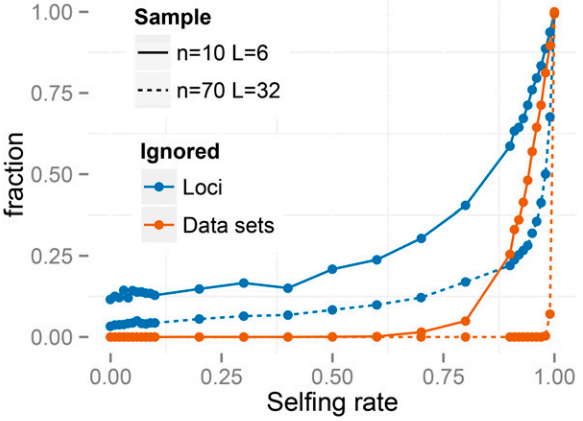


Figure 3 Fraction of loci and data sets that are ignored by RMES.

expectation $s^*/(2-s^*)$ (Wright 1921; Haldane 1924) and solving for s^* :

$$\widehat{s^*} = \frac{2\widehat{F_{\mathrm{IS}}}}{1+\widehat{F_{\mathrm{IS}}}}. \tag{27}$$

In accommodating multiple loci, this estimate incorporates a multilocus estimate for $\widehat{F_{\mathrm{IS}}}$ (*Appendix B*) but, unlike those generated by our Bayesian method and RMES, does not use identity disequilibrium across loci within individuals to infer the number of generations since the most recent outcross event in their ancestry. As our primary purpose in examining the F_{IS}-based estimate (Equation 27) is to provide a baseline for the results of those likelihood-based methods, we have not attempted to develop an index of error or uncertainty for it.

Accuracy

To assess relative accuracy of estimates of the uniparental proportion s^*, we determine the bias and root-mean-square error of the three methods by averaging over 10^4 data sets (10^2 independent samples from each of 10^2 independent simulations for each assigned s^*). In contrast with the point estimates of s^* produced by RMES, our Bayesian method generates a posterior distribution. To facilitate comparison, we reduce our estimate to a single value, the mode of the posterior distribution of s^*, with the caveat that the median and mean may show different qualitative behavior (see File S1).

Figure 2 indicates that our method, RMES, and even the F_{IS}-based estimate (Equation 27) provide estimates of the uniparental proportion s^* that show little bias over most of its range. RMES differs from the other two methods in showing a steep rise in both bias and root-mean-square (RMS) error for high values of s^*, with the change point occurring at lower values of the uniparental proportion s^* for the small-sample regime ($n=10, L=6$). A likely contributing factor to the increased error shown by RMES under high values of s^* is its default assumption that the number of generations in the ancestry of any individual does not exceed 20. Violations of this assumption arise more often under high values of s^*, possibly promoting underestimation of the uniparental proportion. Further, RMES discards data at loci at which no heterozygotes are observed and terminates analysis altogether if the number of loci drops below 2. RMES treats all loci with zero heterozygosity (Equation 1) as uninformative, even if multiple alleles are observed. In contrast, our full-likelihood method uses data from all loci, with polymorphic loci in the absence of heterozygotes providing strong evidence of high rates of selfing (rather than low rates of mutation). Under the large sampling regime ($n=70$, $L=32$), RMES discards on average 50% of the loci for true s^* values exceeding 0.94, with $<$ 10% of data sets unanalyzable ($<$2 informative loci) even at $s^*=0.99$ (Figure 3). Under the $n=10$, $L=6$ regime, RMES discards on average 50% of loci for true s^* values exceeding 0.85, with ~50% of data sets unanalyzable under $s^* \geq 0.94$.

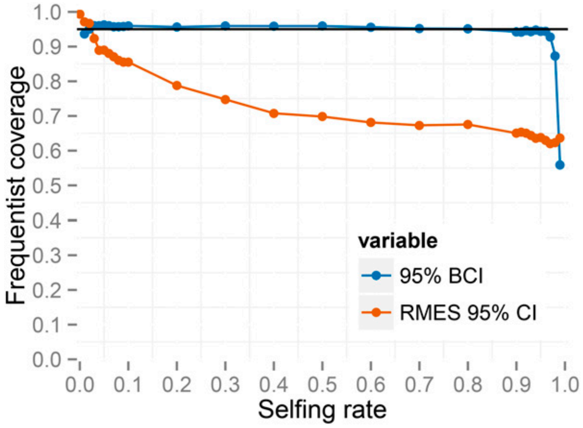


Figure 4 Frequentist coverage at each level of s^* for 95% intervals from RMES and the method based on the full likelihood under the large sampling regime ($n=70$, $L=32$). RMES intervals are 95% confidence intervals computed via profile likelihood. Full-likelihood intervals are 95% highest posterior density Bayesian credible intervals.

Coverage

We determine the fraction of data sets for which the confidence interval (C.I.) generated by RMES and the Bayesian credible interval (BCI) generated by our method contain the true value of the uniparental proportion s^*. This measure of coverage is a frequentist notion, as it treats each true value of s^* separately. A 95% C.I. should contain the truth 95% of the time for each specific value of s^*. However, a 95% BCI is not expected to have 95% coverage at each value of s^*, but rather 95% coverage averaged over values of s^* sampled from the prior. Of the various ways to determine a BCI for a given posterior distribution, we choose to report the highest posterior density BCI (rather than the central BCI, for example).

Figure 4 indicates that coverage of the 95% C.I.'s produced by RMES is consistently <95% across all true s^* values under the large sampling regime ($n=70$, $L=32$). Coverage appears to decline as s^* increases, dropping from 86% for $s^*=0.1$ to 64% for $s^*=0.99$. In contrast, the 95% BCIs have

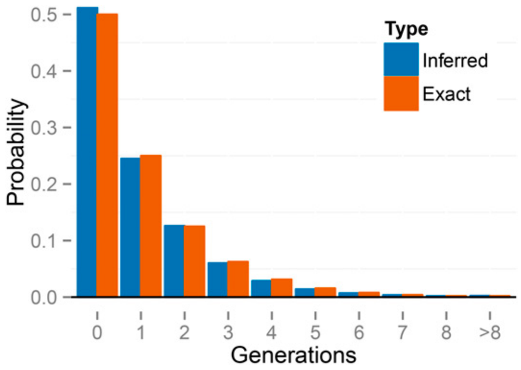


Figure 5 Exact distribution of selfing times under $s^* = 0.5$ compared to the posterior distribution averaged across individuals and across data sets.

slightly >95% frequentist coverage for each value of s^*, except for s^* values very close to the extremes (0 and 1). Under very high rates of inbreeding ($s^* \approx 1$), an assumption (Equation 4) of our underlying model (random outcrossing occurs on a timescale much shorter than the timescales of mutation and coalescence) is likely violated. We observed similar behavior under nominal coverage levels ranging from 0.5 to 0.99 (File S1).

Number of consecutive generations of selfing

To check the accuracy of our reconstructed generations of selfing, we examine the posterior distributions of selfing times $\{T_k\}$ for $s^* = 0.5$ under the large sampling regime $(n = 70,\ L = 32)$. We average posterior distributions for selfing times across 100 simulated data sets and across individuals $k = 1 \ldots 70$ within each simulated data set. We then compare these averages based on the simulated data with the exact distribution of selfing times across individuals (Figure 5). The pooled posterior distribution closely matches the exact distribution. This simple check suggests that our method correctly infers the true posterior distribution of selfing times for each sampled individual.

Analysis of Microsatellite Data from Natural Populations

To illustrate the features of our method, we apply it to existing microsatellite data sets from natural populations of a self-fertilizing vertebrate and a plant. We note that the infinite-alleles model of mutation may fail to capture features of mutation processes of microsatellites.

Self-fertilizing vertebrate

Our analysis of data from the killifish *K. marmoratus* (Mackiewicz *et al.* 2006; Tatarenkov *et al.* 2012) incorporates genotypes from 32 microsatellite loci as well as information on the observed fraction of males. Our method jointly estimates the proportion of males in the population (p_m) together with rates of locus-specific mutation (θ^*) and the uniparental proportion (s_L). We apply the method to two populations, which show highly divergent rates of inbreeding.

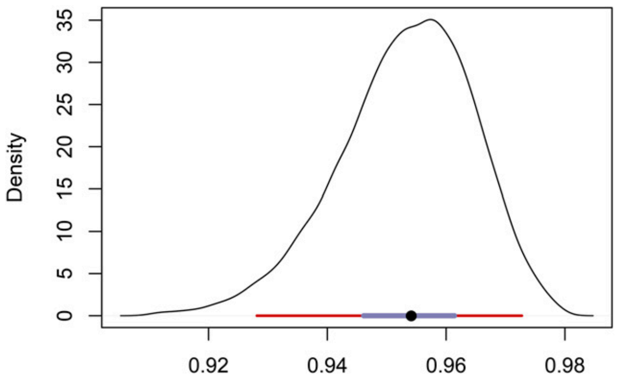


Figure 6 Posterior distribution of the uniparental proportion s_L for the BP population. The median is indicated by a black circle, with a maroon bar for the 95% BCI and a slate-colored bar for the 50% BCI.

Parameter estimation: Our analysis uses an expanded-likelihood expression, which directly incorporates the observation of n_m males among n_{total} zygotes,

$$\begin{aligned}&\Pr\left(\mathbf{X}, \mathbf{I}, \mathbf{T}, n_m | s^*, \mathbf{\Theta}^*, p_m, n_{total}\right)\\&\quad = \Pr\left(\mathbf{X}, \mathbf{I}, \mathbf{T} | s^*, \mathbf{\Theta}^*\right) \cdot \Pr\left(n_m | p_m, n_{total}\right),\end{aligned}$$

in which

$$n_m \sim \text{Binomial}\left(n_{total},\ p_m\right), \tag{28}$$

for p_m (Equation 8a) the fraction of males among reproductives, under the assumption that the sex ratio among observed individuals corresponds to the sex ratio among reproductives. The likelihood expression reflects that s^* and $\mathbf{\Theta}^*$ are sufficient to account for $\mathbf{X}$, $\mathbf{I}$, and $\mathbf{T}$, which are independent of n_m, n_{total}, and p_m.

In the absence of direct information regarding the existence or intensity of inbreeding depression, we impose the constraint $\tau \equiv 1$, which permits estimation of the uniparental proportion s_L under a uniform prior:

$$s^* \sim \text{Uniform}(0,\ 1).$$

Low outcrossing rate: We applied our method to the BP data set described by Tatarenkov *et al.* (2012). This data set comprises a total of 70 individuals, collected in 2007, 2010, and 2011 from the Big Pine location in the Florida Keys.

Tatarenkov *et al.* (2012) report 2 males among the 201 individuals collected from various locations in the Florida Keys during this period, consistent with other estimates of ~1% (*e.g.*, Turner *et al.* 1992). Drawing on the long-term experience of the Tatarenkov–Avise laboratory, we assume observation of $n_m = 20$ males of $n_{total} = 2000$ individuals in (28). Our purpose here is to illustrate the application of the method, with researchers using the software for primary research encouraged to substitute actual numbers. Our analysis for the BP population generates a posterior distribution for the fraction of males in the population (p_m) with a posterior median of 0.01 and a 95% BCI of $(0.0062,\ 0.015)$.

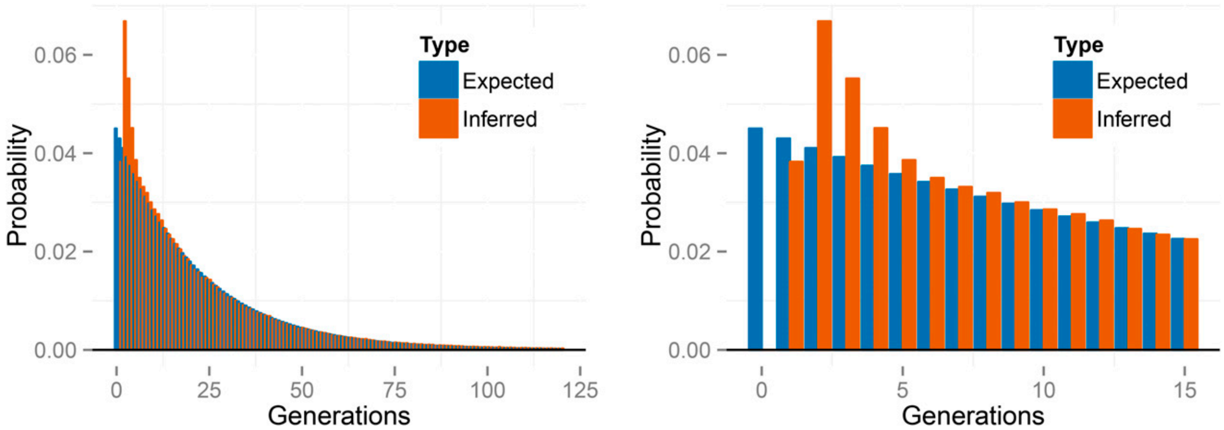


Figure 7 Empirical distribution of number of generations since the most recent outcross event (*T*) across individuals for the BP population of *K. marmoratus*, averaged across posterior samples. The right panel is constructed by zooming in on the panel on the left. "Expected" probabilities represent the proportion of individuals with the indicated number of selfing generations expected under the median uniparental proportion s_L. "Inferred" probabilities represent proportions inferred across individuals in the sample. The first inferred bar with positive probability corresponds to $T = 1$.

Our estimates of mutation rates (θ^*) indicate substantial variation among loci, with the median ranging over an order of magnitude (~0.5–5.0) (Figure S8). The distribution of mutation rates across loci appears to be multimodal, with many loci having a relatively low rate and some having larger rates.

Figure 6 shows the posterior distribution of uniparental proportion s_L, with a median of 0.95 and a 95% BCI of $(0.93,\ 0.97)$. This estimate appears to be somewhat lower than the F_{IS}-based estimate (Equation 27) of 0.97 and slightly higher than the RMES estimate of 0.94, which has a 95% C.I. of $(0.91,\ 0.96)$. We note that RMES discarded from the analysis data from the 9 loci (of 32) that showed no heterozygosity, even though 7 of the 9 were polymorphic in the sample.

Our method estimates the latent variables $\{T_1, T_2, \ldots, T_n\}$ (Equation 13), representing the number of generations since the most recent outcross event in the ancestry of each individual (Figure S6). Figure 7 shows the empirical distribution of the time since outcrossing across individuals, averaged over posterior uncertainty, indicating a complete absence of biparental individuals (0 generations of selfing). Because we expect that a sample of size 70 would include at least some biparental individuals under the inferred uniparental proportion ($s_L \approx 0.95$), this finding suggests that any biparental individuals that may exist in the sample show lower heterozygosity than expected from the observed level of genetic variation. This deficiency suggests that an extended model that accommodates biparental inbreeding or population subdivision may account for the data better than the present model, which allows only selfing and random outcrossing.

Higher outcrossing rate: We apply the three methods to the sample collected in 2005 from Twin Cays, Belize (TC05) (Mackiewicz *et al.* 2006). Compared to the BP data set, this TC data set shows considerably higher incidence of males and levels of polymorphism and heterozygosity.

We incorporate the observation of 19 males among the 112 individuals collected from Belize in 2005 (Mackiewicz *et al.* 2006) into the likelihood (see Equation 28). Our estimate of the population fraction of males among reproductives (p_m) has a posterior median of 0.17 with a 95% BCI of $(0.11,\ 0.25)$.

Figure S9 indicates that the posterior medians of the locus-specific mutation rates span a wide range (~0.5–23). Two loci appear to exhibit mutation rates substantially higher than those of other loci, both of which appear to have high rates in the BP population as well (Figure S8). The rank orders of median mutation rates estimated across loci from the two data sets show only diffuse correspondence (Figure S10).

All three methods confirm the inference of Mackiewicz *et al.* (2006) of much lower inbreeding in the TC population relative to the BP population. Our posterior distribution of uniparental proportion s_L has a median and 95% BCI of 0.35 $(0.25,\ 0.45)$ (Figure 8). This median again lies between the F_{IS}-based estimate (Equation 27) of 0.39 and the RMES estimate of 0.33, which has a 95% C.I. of $(0.30,\ 0.36)$. In this case, RMES excluded from the analysis only a single locus, which was monomorphic in the sample.

Figure 9 shows the inferred distribution of the number of generations since the most recent outcross event (T) across individuals, averaged over posterior uncertainty. In contrast to the BP population, the distribution of time since the most recent outcross event in the TC population appears to conform to the distribution expected under the inferred uniparental proportion (s_L), including a high fraction of biparental individuals ($T_k = 0$). Figure S7 presents the posterior distribution of the number of consecutive generations of selfing in the immediate ancestry of each individual.

Gynodioecious plant

We next examine data from *Schiedea salicaria*, a gynodioecious member of the carnation family endemic to the Hawaiian islands. We analyzed genotypes at nine microsatellite loci from 25 *S. salicaria* individuals collected from west Maui and identified by Wallace *et al.* (2011) as nonhybrids. We use this analysis to illustrate the incorporation of data in addition to the genotypic scores.

Parameter estimation: Campbell *et al.* (2010) reported a 12% proportion of females ($n_{\mathrm{f}} = 27$ females among $n_{\mathrm{total}} = 221$ individuals). As for *Kryptolebias* (Equation 28), we model this information by

$$n_{\mathrm{f}} \sim \mathrm{Binomial}\left(n_{\mathrm{total}}, p_{\mathrm{f}}\right), \tag{29}$$

obtaining estimates from an extended-likelihood function corresponding to the product of $\Pr(n_{\mathrm{f}} | n_{\mathrm{total}},\ p_{\mathrm{f}})$ and the likelihood of the genetic data.

Our analysis assumes equal seed set by females and hermaphrodites ($\sigma \equiv 1$), consistent with empirical results (Weller and Sakai 2005). In addition, we use results of experimental studies of inbreeding depression to develop an informative prior distribution for τ,

$$\tau \sim \mathrm{Beta}\left(2,\ 8\right), \tag{30}$$

the mean of which (0.2) is consistent with the results of greenhouse experiments reported by Sakai *et al.* (1989). We retain a uniform prior for the proportion of seeds of hermaphrodites set by self-pollen ($\tilde{s}$).

Results: Table 1 presents posterior medians and 95% BCIs for the proportion of uniparentals among reproductives (s_{G}), the proportion of seeds of hermaphrodites set by self-pollen ($\tilde{s}$), the viability of uniparental relative to biparental offspring (τ), the proportion of females among reproductives (p_{f}), and the probability that a random gene derives from a female parent $[(1 - s_{\mathrm{G}})F/2]$. Our full analysis incorporates genetic data (G), observations on the sex ratio (F), and an informative prior (I) for the relative viability of uniparentals (τ) based on results of manipulative experiments (Sakai *et al.* 1989). Each row represents an analysis that includes (Y) or excludes (N) information of type G, F, or I. Comparison of the YYY and NYY rows indicates that inclusion of the genetic data more than doubles the posterior median of s^* (from 0.112 to 0.247) and shrinks the credible interval. Comparison of the YYY and YYN rows indicates that information about the level of inbreeding depression increases the posterior median of the collective contribution of females to the gene pool $[(1 - s_{\mathrm{G}})F/2]$, bringing it closer to the proportion of females (p_{f}), with equality (10b) implying maximization of relative effective number E_{G} (Equation 9d).

Analysis in the absence of data (NNN, bottom row of Table 1) provides a prior estimate for the proportion of uniparentals (s_{G}) of $0.0844\ (0.000797,\ 0.643)$. While the proportion of seeds set by self-pollen ($\tilde{s}$) and the relative viability of uniparental offspring (τ) have uniform priors, the induced prior on composite parameter s_{G} departs from uniform on $(0, 1)$.

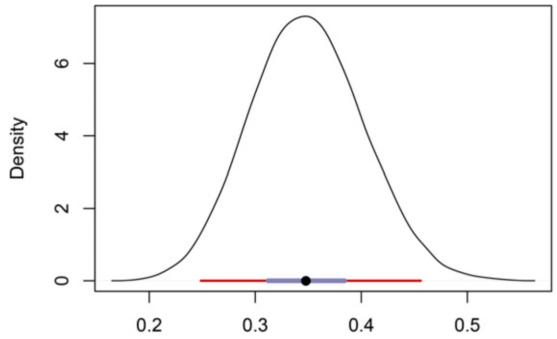


Figure 8 Posterior distribution of the uniparental proportion s_{L} for the TC population. Also shown are the 95% BCI (maroon), 50% BCI (slate color), and median (black circle).

In the absence of information about $\tilde{s}$ and τ, we recommend that researchers use the pure hermaphrodite model (Equation 6) with τ assigned as unity so that s_{G} will be estimated under a uniform prior. We adopt this approach to compare our method to RMES, which uses only the genotype counts. Our estimate of the uniparental proportion s_{G} [median 0.287, 95% BCI $(0.110,\ 0.478)$] is similar to the estimate using all information (YYY in Table 1) and in line with the F_{IS}-based estimate (Equation 27) of $s_{\mathrm{G}} = 0.33$. In contrast, RMES gave an estimate of $s_{\mathrm{G}} = 0$ [95% CI $(0,\ 0.15)$], even though it excluded none of the loci. Application of our gynodioecy model to the genotypic counts with or without additional information (YYY, YYN, YNY, or YNN in Table 1) produces estimates of the selfing rate for which the 95% BCIs exclude zero. This unexpected estimate of RMES stands in opposition to previous work supporting the presence of selfing in this population of *S. salicaria* (Wallace *et al.* 2011).

Figure 10 presents the inferred distribution across individuals of the number of generations since the most recent outcross event T (Equation 15), averaged over posterior uncertainty, using all data (YYY). In contrast with the analysis of the *K. marmoratus* BP population (Figure 7), the distribution appears to be consistent with the inferred uniparental proportion s_{G}.

We include additional results obtained using all data (YYY) in File S1. Figure S11 presents posterior distributions of all basic parameters of the gynodioecy model (Equation 9). Unlike the *K. marmoratus* data sets, the *S. salicaria* data set does not appear to provide substantial evidence for large differences in locus-specific mutation rates across loci (Figure S13). Figure S12 presents the posterior distribution of the number of consecutive generations of selfing in the immediate ancestry of each individual.

Discussion

We introduce a model-based Bayesian method for the inference of the rate of self-fertilization and other aspects of mating systems. Designed to accommodate arbitrary numbers

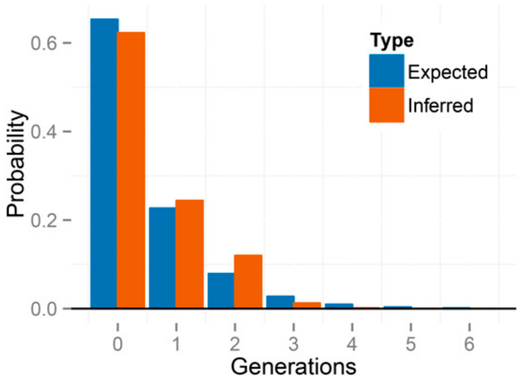


Figure 9 Empirical distribution of selfing times *T* across individuals, for *K. marmoratus* (population TC). The histogram is averaged across posterior samples.

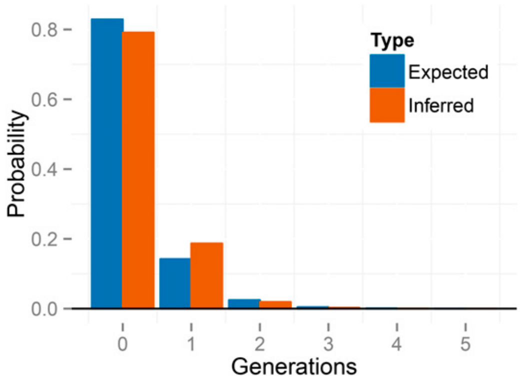


Figure 10 Empirical distribution of selfing times *T* across individuals, for *S. salicaria*. The histogram is averaged across posterior samples.

of loci, it uses the ESF to determine likelihoods in a computationally efficient manner from frequency spectra of genotypes observed at multiple unlinked sites throughout the genome. Our MCMC sampler explicitly incorporates the full set of parameters for each mating system considered (pure hermaphroditism, *Kryptolebias*, and gynodioecy). This construction permits incorporation of information in addition to genetic data, affording insight into components of the evolutionary process beyond the estimation of selfing rates alone.

Components of inference

Locus-specific mutation rates: Our method permits variation among loci in the rate of mutation (Equation 3) by using the DPP to determine the number of rate classes, the mutation rate of each class, and the class to which each locus belongs. Our DPP adopts a conservative approach, creating a new rate class only if the data demand it. Under the DPP, loci belonging to the same group have identical mutation rates. This approach might be generalized, for example, by using a Dirichlet process mixture to allow variation in mutation rate among loci within a rate class.

Joint inference of mutation and inbreeding rates: For the infinite-alleles model of mutation, the ESF (Ewens 1972) provides the probability of any allele frequency spectrum (AFS) observed at a locus in a sample derived from a panmictic population. Under partial self-fertilization, the ESF provides the probability of an AFS observed among genes, each sampled from a distinct individual. For such genic (as opposed to genotypic) samples, the coalescence process under inbreeding is identical to the standard coalescence process, but with a rescaling of time (Fu 1997; Nordborg and Donnelly 1997). Accordingly, genic samples may serve as the basis for the estimation of the single parameter of the ESF, the scaled mutation rate θ^* (Equation 5), but not the rate of inbreeding apart from the scaled mutation rate.

Our method uses the information in a genotypic sample, the genotype frequency spectrum, to infer both the uniparental proportion s^* and the scaled mutation rate θ^*. Our sampler reconstructs the genealogical history of a sample of diploid genotypes only to the point of the most recent random-outcross event of each individual, with the number of consecutive generations of inbreeding in the immediate ancestry of a given individual (T_k for individual k) corresponding to a latent variable in our Bayesian inference framework. Invocation of the ESF beyond the point at which all lineages reside in separate individuals obviates the necessity of further genealogical reconstruction. As a consequence, our method may be better able to accommodate genome-scale magnitudes of observed loci (L).

Identity disequilibrium (Cockerham and Weir 1968), the correlation in heterozygosity across loci within individuals, reflects that all loci within an individual experience the most recent random-outcross event at the same time, irrespective of physical linkage. The heterozygosity profile of individual k provides information about T_k (Equation 15), which in turn reflects the uniparental proportion s^*. Observation of multiple individuals provides a basis for inference of both the uniparental proportion s^* and the scaled mutation rate θ^*.

Estimation of the selfing rate

Accuracy and uncertainty: Enjalbert and David (2000) and David *et al.* (2007) base estimates of selfing rate on the distribution of numbers of heterozygous loci. Both methods strip genotype information from the data, distinguishing between only homozygotes and heterozygotes, irrespective of the alleles involved. Loci lacking heterozygotes altogether (even if polymorphic) are removed from the analysis as uninformative about the magnitude of departure from Hardy–Weinberg proportions (Figure 3). As the observation of polymorphic loci with low heterozygosity provides strong evidence of inbreeding, exclusion of such loci by RMES (David *et al.* 2007) may contribute to its loss of accuracy for high rates of selfing (Figure 2).

Our method derives information from all loci. Like most coalescence-based models, it accounts for the level of variation as well as the way in which variation is partitioned within the sample. Even a locus monomorphic within a sample provides information about the age of the most recent

Table 1 Parameter estimates for different amounts of data

G	F	I	s_G	$\tilde{s}$	τ	p_f	$(1-s_G)F/2$
Y	Y	Y	0.247 (0.0791, 0.444)	0.695 (0.299, 0.971)	0.215 (0.0597, 0.529)	0.125 (0.0849, 0.173)	0.118 (0.054, 0.258)
Y	Y	N	0.267 (0.0951, 0.469)	0.497 (0.187, 0.93)	0.507 (0.082, 0.973)	0.125 (0.0851, 0.174)	0.0808 (0.0398, 0.191)
Y	N	Y	0.213 (0.045, 0.402)	0.742 (0.379, 1.00)	0.252 (0.0488, 0.529)	0.244 (0.00, 0.613)	0.218 (0.0, 0.403)
Y	N	N	0.243 (0.0608, 0.429)	0.628 (0.268, 0.999)	0.611 (0.167, 1.00)	0.354 (0.00, 0.072)	0.223 (0.00, 0.394)
N	Y	Y	0.112 (0.0026, 0.588)	0.496 (0.0252, 0.974)	0.183 (0.0277, 0.513)	0.125 (0.0847, 0.173)	0.0956 (0.0427, 0.218)
N	Y	N	0.231 (0.00391, 0.776)	0.504 (0.025, 0.973)	0.493 (0.0257, 0.975)	0.125 (0.0847, 0.173)	0.0778 (0.0392, 0.172)
N	N	Y	0.0376 (0.00, 0.318)	0.492 (0.0122, 0.957)	0.0.185 (0.00917, 0.462)	0.483 (0.00, 0.946)	0.314 (0.0361, 0.500)
N	N	N	0.0844 (0.000, 0.643)	0.497 (0.0244, 0.975)	0.494 (0.0252, 0.975)	0.479 (0.0245, 0.972)	0.289 (0.0313, 0.5)

Estimates are given by a posterior median and a 95% BCI. Each row represents an analysis that includes (Y) or excludes (N) information, including genotype frequency data (G), counts of females (F), and replacement of the Uniform(0,1) prior on τ by an informative prior (I).

common ancestor of the observed sequences, a property that was not widely appreciated prior to analyses of the absence of variation in a sample of human Y chromosomes (Dorit *et al.* 1995; Fu and Li 1996).

Both RMES and our method invoke independence of genealogical histories of unlinked loci, conditional on the time since the most recent outcrossing event. RMES seeks to approximate the likelihood by summing over the distribution of time since the most recent outcross event, but truncates the infinite sum at 20 generations. The increased error exhibited by RMES under high rates of inbreeding may reflect that the likelihood has a substantial mass beyond the truncation point in such cases. Our method explicitly estimates the latent variable of time since the most recent outcross for each individual (Equation 13). This quantity ranges over the nonnegative integers, but values assigned to individuals are explored by the MCMC according to their effects on the likelihood.

Estimates of the proportion of uniparental individuals s^* (Equation 4) produced by our method appear to show greater accuracy than RMES over much of the parameter range (Figure 2). Even so, we note that all methods considered here provide fair estimates of the selfing rate, including the F_{IS}-based method (Equation 27) that uses only the single-locus departures from Hardy–Weinberg proportions and not identity disequilibrium. However, our Bayesian method appears to provide a more accurate assessment of uncertainty than does the maximum-likelihood method RMES: our BCIs have good frequentist coverage properties (Figure S5), while the C.I.'s reported by RMES appear to perform less well (Figure 4).

Identifiability: In an analysis based solely on the genotype frequency spectrum observed in a sample, the likelihood depends on just two composite parameters: the probability that a random individual is uniparental (s^*) and the scaled rates of mutation $\mathbf{\Theta}^*$ (Equation 19) across loci. Even so, our MCMC implementation updates the full set of basic parameters, with likelihoods determined from the implied values of s^* and $\mathbf{\Theta}^*$.

Any model for which the parameter set $\mathbf{\Psi}$ (Equation 23) comprises more than one parameter is not fully identifiable from the genetic data alone. In the pure hermaphroditism model (Equation 6), for example, basic parameters $\tilde{s}$ (fraction of fertilizations by selfing) and τ (relative viability of uniparental offspring) are nonidentifiable: any assignments that determine the same values of composite parameters s^* and $\mathbf{\Theta}^*$ have the same likelihood.

For each basic parameter in $\mathbf{\Psi}$ beyond one, identifiability requires incorporation of additional information beyond the genetic data. A full treatment of such information requires expansion of the likelihood function to encompass an explicit model of the new information. For example, the *Kryptolebias* model (Equation 7) comprises three basic parameters, including p_{m} (Equation 8a), the frequency of males among reproductives. In our analysis of microsatellite data from the killifish *K. marmoratus* (Mackiewicz *et al.* 2006; Tatarenkov *et al.* 2012), the expanded-likelihood function corresponds to the product of the probability of the genetic data and the probability of the number of males observed among a total number of individuals (Equation 28). In a similar manner, our analysis of the data set from *S. salicaria* (Wallace *et al.* 2011) uses an extended-likelihood function that models the observed number of females as a binomial random variable (Equation 29), permitting estimation of the frequency of females among reproductives (p_{f}).

Nonidentifiable parameters can also be estimated through the incorporation of informative priors. Because identifiability is defined in terms of the likelihood, which is unaffected by priors, such parameters remain nonidentifiable. Even so, informative priors assist in their estimation through Bayesian approaches, which do not require parameters to be identifiable. Our analysis of the *Schiedea* data draws on experimental evidence in addition to the genotype counts to justify the assumption of equal seed set by females and hermaphrodites ($\sigma \equiv 1$) (Weller and Sakai 2005) and to develop an informative prior for τ (Equation 30) (Sakai *et al.* 1989).

Guidance for applying the method: Our present implementation of the method introduced here includes default priors for the basic parameters, with users encouraged to specify priors appropriate for their systems. For example, a biologically motivated prior for the relative viability of uniparentals (τ) might favor weak selection ($\tau \approx 1$) or inbreeding depression of an intensity sufficient to maintain selfing ($\tau \geq 1/2$).

In the *Kryptolebias* model (Equation 7), comprising basic parameters $\tilde{s}$ (proportion of eggs self-fertilized by hermaphrodites),

τ (relative viability of uniparentals), and p_m (proportion of males among reproductives), p_m together with s_L determines the scaling of time (Equation 7c), which depends on $\tilde{s}$ and τ only through s_L. In the absence of information regarding $\tilde{s}$ and τ, we recommend assigning $\tau \equiv 1$, which permits estimation of s_L under the default uniform prior or a user-specified prior. This assignment produces estimates that are simply agnostic concerning the relative influence of $\tilde{s}$ and τ in determining s_L.

In the four-parameter gynodioecy model (Equation 26), however, the scaling of time (Equation 9d) depends not only on s_G (the proportion of uniparentals) and p_f (the proportion of females among reproductives), but also on F (the proportion of biparental offspring that have a female parent). Because F (Equation 9b) depends on $\tilde{s}$ (the proportion of seeds set by self-pollen), information about τ affects inference of all basic parameters. In the absence of information about the intensity of inbreeding depression, we recommend using the pure hermaphroditism model (Equation 24) under the assignment $\tau \equiv 1$, which permits estimation of the uniparental proportion s^* under a uniform prior.

Beyond estimation of the selfing rate

Unlike the other methods considered here, our method provides a framework for the incorporation of information in addition to counts of diploid genotypes and the inference of a number of aspects of the mating system beyond the selfing rate.

Time since the most recent outcross: Our method incorporates as a latent variable T_k (Equation 13), the number of generations since the most recent outcross event in the immediate ancestry of individual k, and provides posterior distributions for this quantity.

This aspect of the mating system is of biological interest in itself and also affords insight into the suitability of the underlying model. Pooling such estimates of times since the most recent outcross over individuals produces an empirical distribution of the number of consecutive generations of selfing. Under the assumption of a single population-wide rate of self-fertilization, we expect selfing time to have a geometric distribution with parameters corresponding to the estimated selfing rate. Empirical distributions of the estimated number of generations since the last outcross appear consistent with this expectation for the data sets derived from the TC population of *K. marmoratus* (Figure 9) and from *Schiedea* (Figure 10). In contrast, the empirical distribution for the highly inbred BP population of *K. marmoratus* (Figure 7) shows an absence of individuals formed by random outcrossing ($T = 0$).

That our method accurately estimates T from simulated data (Figure 5) argues against attributing the inferred deficiency of biparental individuals in the BP data set to an artifact of the method. Rather, the deficiency may indicate a departure from the underlying model, which assumes reproduction only through self-fertilization or random outcrossing. In particular, biparental inbreeding as well as selfing may reduce the fraction of individuals formed by random outcrossing. Misscoring of heterozygotes as homozygotes due to null alleles or other factors, a possibility directly addressed by RMES (David *et al.* 2007) but not by our method, may also in principle contribute to the apparent deficiency of individuals formed by random outcrossing.

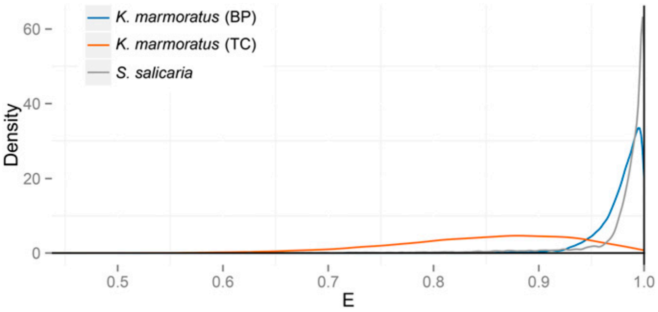


Figure 11 Posterior distributions of relative effective number E (Equation 3) for data sets derived from *K. marmoratus* (BP and TC populations) and *S. salicaria*.

Relative effective number: Incorporation of additional information, either through extension of the likelihood or through informative priors, permits inference not only of the basic parameters but also of functions of the basic parameters. For example, Table 1 includes estimates of the proportion of seeds of hermaphrodites set by self-pollen ($\tilde{s}$) and the probability that a random gene derives from a female parent $[(1 - s_G)F/2]$ in gynodioecious *S. salicaria*. We are not aware of other studies in which these quantities have been inferred from the pattern of neutral genetic variation observed in a random sample.

Among the most biologically significant functions to which this approach affords access is relative effective number E (Equation 3), a fundamental component of the reproductive value of the sexes (Fisher 1958). We denote the probability that a pair of genes, randomly drawn from distinct individuals, derive from the same parent in the preceding generation as the rate of parent sharing ($1/N^*$). Its inverse (N^*) corresponds to the inbreeding effective size of Crow and Denniston (1988). Relative effective number E is the ratio of N^* to the total number of reproductive individuals. For example, in the absence of inbreeding ($s^* = 0$), N^* in our gynodioecy model (Equation 9) corresponds to Wright's (1969) harmonic mean expression for effective population size and E to the ratio of N^* and $N_f + N_h$, the total number of reproductive females and hermaphrodites. In general ($s^* \geq 0$), relative effective size E reflects reductions in effective size due to inbreeding in addition to differences in numbers of the sexual forms.

Figure 11 presents posterior distributions of E for the three data sets explored here. These results suggest that relative effective number E in each of the natural populations surveyed lies close to its maximum of unity, with the effective number defined through the rate of parent sharing approaching the total number of reproductives. Our estimates suggest that maximization of relative effective number would occur under a slight increase in the frequency of males p_m (Equation 8b) in both *K. marmoratus* populations and a very slight decrease in the frequency of females p_f (Equation 10b) in the *S. salicaria* population.

Acknowledgments

We thank the Reviewers, Associate Editor Rasmus Nielsen, and Senior Editor John Wakeley for valuable comments and suggestions. This project was initiated during a sabbatical visit of M.K.U. to the Department of Ecology and Evolutionary Biology at the University of California at Irvine. We thank Francisco J. Ayala for exceedingly gracious hospitality and Diane R. Campbell and all members of the Department for stimulating interactions. We thank Lisa E. Wallace for making available to us microsatellite data from *S. salicaria*. Public Health Service grant GM 37841 (to M.K.U.) provided partial funding for this research.

Literature Cited

Ayres, K. L., and D. J. Balding, 1998 Measuring departures from Hardy-Weinberg: a Markov chain Monte Carlo method for estimating the inbreeding coefficient. Heredity 80: 769–777.

Campbell, D. R., S. G. Weller, A. K. Sakai, T. M. Culley, P. N. Dang *et al.*, 2010 Genetic variation and covariation in floral allocation of two species of *Schiedea* with contrasting levels of sexual dimorphism. Evolution 65: 757–770.

Clegg, M. T., 1980 Measuring plant mating systems. Bioscience 30: 814–818.

Cockerham, C. C., and B. S. Weir, 1968 Sib mating with two linked loci. Genetics 60: 629–640.

Crow, J. F., and C. Denniston, 1988 Inbreeding and variance effective population numbers. Evolution 42: 482–495.

David, P., B. Pujol, F. Viard, V. Castella, and J. Goudet, 2007 Reliable selfing rate estimates from imperfect population genetic data. Mol. Ecol. 16: 2474–2487.

Dorit, R. L., H. Akashi, and W. Gilbert, 1995 Absence of polymorphism at the ZFY locus on the human Y chromosome. Science 286: 1183–1185.

Enjalbert, J., and J. L. David, 2000 Inferring recent outcrossing rates using multilocus individual heterozygosity: application to evolving wheat populations. Genetics 156: 1973–1982.

Ewens, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Popul. Biol. 3: 87–112.

Fisher, R. A., 1958 *The Genetical Theory of Natural Selection*, Ed. 2. Dover, New York.

Fu, Y.-X., 1997 Coalescent theory for a partially selfing population. Genetics 146: 1489–1499.

Fu, Y.-X., and W.-H. Li, 1996 Absence of polymorphism at the ZFY locus on the human Y chromosome. Science 272: 1356–1357.

Furness, A. I., A. Tatarenkov, and J. C. Avise, 2015 A genetic test for whether pairs of hermaphrodites can cross-fertilize in a selfing killifish. J. Hered. DOI:10.1093/jhered/esv077.

Gao, H., S. Williamson, and C. D. Bustamante, 2007 A Markov chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. Genetics 176: 1635–1651.

Griffiths, R. C., and S. Lessard, 2005 Ewens' sampling formula and related formulae: combinatorial proofs, extensions to variable population size and applications to ages of alleles. Theor. Popul. Biol. 68: 167–177.

Haldane, J., 1924 A mathematical theory of natural and artificial selection. Part ii. The influence of partial self-fertilization, inbreeding, assortative mating, and selective fertilization on the composition of Mendelian populations, and on natural selection. Biol. Rev. Camb. Philos. Soc. 1: 158–163.

Hill, W. G., H. A. Babiker, L. C. Ranford-Cartwright, and D. Walliker, 1995 Estimation of inbreeding coefficients from genotypic data on multiple alleles, and application to estimation of clonality in malaria parasites. Genet. Res. 65: 53–61.

Karlin, S., and J. McGregor, 1972 Addendum to a paper of W. Ewens. Theor. Popul. Biol. 3: 113–116.

Liu, J. S., 2001 *Monte Carlo Strategies in Scientific Computing*. Springer, New York.

Mackiewicz, M., A. Tatarenkov, D. S. Taylor, B. J. Turner, and J. C. Avise, 2006 Extensive outcrossing and androdioecy in a vertebrate species that otherwise reproduces as a self-fertilizing hermaphrodite. Proc. Natl. Acad. Sci. USA 103: 9924–9928.

Neal, R. M., 2000 Markov chain sampling methods for Dirichlet process mixture models. J. Comput. Graph. Stat. 9: 249–265.

Nordborg, M., and P. Donnelly, 1997 The coalescent process with selfing. Genetics 146: 1185–1195.

Pritchard, J. K., M. Stephens, and P. Donnelly, 2000 Inference of population structure using multilocus genotype data. Genetics 155: 945–959.

Ritland, K., 2002 Extensions of models for the estimation of mating systems using n independent loci. Heredity 88: 221–228.

Sakai, A. K., K. Karoly, and S. G. Weller, 1989 Inbreeding depression in *Schiedea globosa* and *S. salicaria* (Caryophyllaceae), subdioecious and gynodioecious Hawaiian species. Am. J. Bot. 76: 437–444.

Tatarenkov, A., R. L. Earley, D. S. Taylor, and J. C. Avise, 2012 Microevolutionary distribution of isogenicity in a self-fertilizing fish (*Kryptolebias marmoratus*) in the Florida Keys. Integr. Comp. Biol. 52: 743–752.

Turner, B. J., W. P. Davis, and D. S. Taylor, 1992 Abundant males in populations of a selfing hermaphrodite fish, *Rivulus marmoratus*, from some Belize cays. J. Fish Biol. 40: 307–310.

Wallace, L. E., T. M. Culley, S. G. Weller, A. K. Sakai, A. Kuenzi *et al.*, 2011 Asymmetrical gene flow in a hybrid zone of Hawaiian Schiedea (Caryophyllaceae) species with contrasting mating systems. PLoS One 6: e24845.

Wang, J., Y. A. El-Kassaby, and K. Ritland, 2012 Estimating selfing rates from reconstructed pedigrees using multilocus genotype data. Mol. Ecol. 21: 100–116.

Weir, B. S., 1996 *Genetic Data Analysis II*. Sinauer Associates, Sunderland, MA.

Weller, S. G., and A. K. Sakai, 2005 Inbreeding and resource allocation in *Schiedea salicaria* (Caryophyllaceae), a gynodioecious species. J. Evol. Biol. 18: 301–308.

Wright, S., 1921 Systems of mating. I, II, III, IV, V. Genetics 6: 111–178.

Wright, S., 1969 *The Theory of Gene Frequencies* (Evolution and the Genetics of Populations, Vol. 2). University of Chicago Press, Chicago.

Communicating editor: R. Nielsen

Appendix A

The Last-Sampled Gene

We present a first-principles derivation (not requiring knowledge of the ESF) of the probability that the last-sampled gene of i genes randomly sampled from distinct individuals represents a novel allele (Equation 21a).

Under the infinite-alleles model of mutation, a single mutation in a lineage suffices to distinguish a new allele. We designate as the focal gene the last-sampled gene and consider the level of the genealogical tree in which its ancestral lineage either receives a mutation or joins the gene tree of the sample at size $(i-1)$. Level ℓ of the entire (i-gene) gene tree corresponds to the segment in which ℓ lineages persist.

The probability that the line of descent of the focal gene terminates in a mutation immediately, in level i of the genealogy, is

$$\frac{u}{iu+\binom{i}{2}\Big/N^*(2-s^*)}=\frac{\theta^*}{i(\theta^*+\ i-1)}.$$

In general, the probability that the lineage of the focal gene terminates on level $\ell > 2$ is

$$\frac{(i-1)u+\binom{i-1}{2}\Big/N^*(2-s^*)}{iu+\binom{i}{2}\Big/N^*(2-s^*)}\,\frac{(i-2)u+\binom{i-2}{2}\Big/N^*(2-s^*)}{(i-1)u+\binom{i-1}{2}\Big/N^*(2-s^*)}$$

$$\cdots\ \frac{lu+\binom{l}{2}\Big/N^*(2-s^*)}{(l+1)u+\binom{l+1}{2}\Big/N^*(2-s^*)}\,\frac{u}{lu+\binom{l}{2}\Big/N^*(2-s^*)}$$

$$=\frac{\theta^*}{i(\theta^*+i-1)}.$$

This expression illustrates the invariance over termination orders noted by Griffiths and Lessard (2005). Summing over all levels, including level 2, for which a mutation in either remaining lineage ensures that the focal gene represents a novel allele, we obtain the overall probability that the last-sampled gene represents a novel allele:

$$\frac{\theta^*(i-2)}{i(\theta^*+i-1)}+\frac{2\theta^*}{i(\theta^*+i-1)}=\frac{\theta^*}{\theta^*+i-1}.$$

Appendix B

Estimators of F_{IS}

We follow Weir (1996) in developing an estimate of the uniparental proportion s^* from F_{IS} alone (Equation 27).

For a single locus, a simple estimator of F_{IS} corresponds to

$$\widehat{F_{\mathrm{IS}}}=1-\frac{O}{E},$$

for O the observed fraction of heterozygotes in the sample and E the expected fraction based on Hardy–Weinberg proportions given the observed allele frequencies. Explicitly, we have

$$\widehat{F_{\mathrm{IS}}}=1-\frac{1-\sum_u\tilde{P}_{uu}}{1-\sum_u\tilde{p}_u^2}=\frac{\left(\sum_u\tilde{P}_{uu}-\tilde{p}_u^2\right)}{1-\sum_u\tilde{p}_u^2},$$

for $\tilde{p}_u$ the frequency of allele u in the sample and $\tilde{P}_{uu}$ the frequency of homozygous genotype uu in the sample. However, this estimator can be substantially biased for small samples, leading to underestimation of F_{IS} (Weir 1996).

To address this bias and accommodate multiple loci, we instead adopt

$$\widehat{F_{\mathrm{IS}}} = \frac{\sum_{l=1}^{L}\left[\sum_{u=1}^{K_l}\left(\tilde{P}_{luu} - \tilde{p}_{lu}^2\right) + \left(1 - \sum_{u=1}^{K_l}\tilde{P}_{luu}\right)\Big/2n\right]}{\sum_{l=1}^{L}\left[\left(1 - \sum_{u=1}^{K_l}\tilde{p}_{lu}^2\right) - \left(1 - \sum_{u=1}^{K_l}\tilde{P}_{luu}\right)\Big/2n\right]}, \tag{B1}$$

for n the number of diploid genotypes observed, L the number of loci, and K_l the number of alleles at locus l. While this estimator is also biased in general, it corresponds to the ratio of unbiased estimators of $F_{\mathrm{IS}} \cdot \sum_l(1 - \sum_u p_{lu}^2)$ and $\sum_l(1 - \sum_u p_{lu}^2)$, in which p_{lu} is the frequency of allele u at locus l in the entire population (Weir 1996). Our analysis of simulated data (*Appendix D*) indicates that this estimator is more accurate than an estimator that simply averages single-locus estimates:

$$\widehat{F_{\mathrm{IS}}} = \frac{1}{L}\sum_{l=1}^{L}\frac{\sum_{u=1}^{K_l}\left(\tilde{P}_{luu} - \tilde{p}_{lu}^2\right) + \left(1 - \sum_{u=1}^{K_l}\tilde{P}_{luu}\right)\Big/2n}{\left(1 - \sum_{u=1}^{K_l}\tilde{p}_{lu}^2\right) - \left(1 - \sum_{u=1}^{K_l}\tilde{P}_{luu}\right)\Big/2n}. \tag{B2}$$

Our F_{IS}-based estimates (Equation 27) incorporate (B1) and not (B2).

Appendix C

Implementation of the MCMC

State space

The state space for the Markov chain of our MCMC sampler includes times across sampled individuals since the last outcross event $\mathbf{T}$ (Equation 13), coalescence events across individuals and loci since that event $\mathbf{I}$ (Equation 14), and model-specific parameters $\boldsymbol{\Psi}$ (Equation 23). The state space also comprises the scaled mutation rates $\boldsymbol{\Theta}^*$ (Equation 19), which are determined by $\mathbf{C}$, a list specifying the mutation rate category C_l for locus $l = 1 \ldots L$, and $\mathbf{Z}$, a list specifying the scaled mutation rate Z_i for category $i = 1 \ldots L + 4$. For example, the scaled mutation rate at locus l corresponds to

$$\theta_l^* = Z_{C_l}. \tag{C1}$$

While the actual number of observed rate categories does not exceed the number of loci (L), expanding the size of the lists to $L + 4$ improves mixing of the MCMC by ensuring that multiple categories are available for the placement of a new, previously unobserved category (see section 6 of Neal 2000). At any given point in the MCMC, the state of the Markov chain corresponds to $(\mathbf{I}, \mathbf{T}, \boldsymbol{\Psi}, \mathbf{C}, \mathbf{Z})$.

Iterations

Each iteration of our MCMC sampler performs multiple updates, with each variable updated at least once per iteration. We recorded the state sampled by the MCMC at each iteration. We assessed convergence by computing for each parameter an effective number of independent samples [effective sample size (Liu 2001)]. Effective numbers for the large sample regime exceeded 500 samples for each parameter on average, while effective number for the small sample regime exceeded 250 samples for each parameter on average. We found that 2000 iterations were more than sufficient to achieve these effective numbers. About 14.5 min were required to complete 2000 iterations for the large sample regime on a Core i3–4030 processor; ~10 sec were required for the small sample regime.

For analyses of simulated data sets, we ran Markov chains for 2000 iterations, discarding the first 200 iterations as burn-in. For analyses of the actual data sets, we ran Markov chains for 100,000 iterations, discarding the first 10,000 iterations as burn-in. Convergence appeared to occur as rapidly for actual data as for simulated data, but we found empirically that the larger number of samples was needed to achieve smooth density plots for the actual data sets.

Transition kernels

Updating of the continuous variables of mutation rates $\{Z_l\}$ (Equation C1) and model-specific parameters $\boldsymbol{\Psi}$ (Equation 23) uses both Metropolis–Hastings (MH) transition kernels and autotuned slice-sampling transition kernels. Updating of the discrete variables $\{C_l\}$ uses a Gibbs transition kernel.

Efficient inference on selfing times through collapsed Metropolis–Hastings

Simple MH proposals that separately update the time since the most recent outcross event (T_k) and coalescence history since that event ($I_{\cdot k}$) lead to extremely poor mixing efficiency. Strong correlations between T_k and $I_{\cdot k}$ cause changes to T_k to be rejected with high probability unless $I_{\cdot k}$ is updated as well. For example, consider proposing a change of T_k from 1 to 0. When $T_k = 1$, on average I_{lk} will be 1 at half of the loci and 0 at the remaining loci. If any of the $I_{lk} = 1$, a move to $T_k = 0$ will always be rejected because the probability of a coalescence event more recently than the most recent outcross event is 0 if the sampled individual is itself a product of outcrossing. To permit acceptance of changes to T_k, we introduce a proposal for T_k that also changes $I_{\cdot k}$.

The scheme starts from the value $T_k = t_k$ and proposes a new value t_k'. In standard MH within Gibbs, we would compute the probability of $T_k = t_k$ and of $T_k = t_k'$ given that all other parameters are unchanged. We modify this MH scheme to compute probabilities without conditioning on the coalescence indicators for individual k. However, the coalescence indicators for other individuals are still held constant. To compute this probability, let J indicate all the coalescence indicators $I_{\cdot y}$ where $y \neq k$. Then

$$\Pr(\mathbf{X}, \mathbf{T}, \mathbf{J}, s, \theta) = \Pr(\mathbf{X}, \mathbf{J} | \mathbf{T}, s, \theta) \Pr(\mathbf{T} | s) \Pr(s) \Pr(\theta).$$

We introduce $\mathbf{I}_{\cdot k}$ by summing over all possible values $\mathbf{i}_{\cdot k}$:

$$\Pr(\mathbf{X}, \mathbf{J} | \mathbf{T}, s, \theta) = \sum_{i_{\cdot k}} \Pr(\mathbf{X}, \mathbf{I}_{\cdot k} = \mathbf{i}_{\cdot k}, \mathbf{J} | \mathbf{T}, s, \theta).$$

Since the i_{lk} for different loci are independent given T_k, we have

$$\begin{aligned}\Pr(\mathbf{X}, \mathbf{J} | \mathbf{T}, s, \theta) &= \sum_{i_{\cdot k}} \prod_{l=1}^{L} \Pr(\mathbf{X}_l,\ I_{lk} = i_{lk}, \mathbf{J}_l | \mathbf{T}, s, \theta) \\ &= \prod_{l=1}^{L} \sum_{i_{lk}} \Pr(\mathbf{X}_l, \mathbf{I}_{lk} = i_{lk}, \mathbf{J}_l | \mathbf{T}, s, \theta).\end{aligned}$$

Therefore, for specific values of $\mathbf{T}$ and $\mathbf{J}$, we can compute the sum over all possible values of $\mathbf{I}_{\cdot k}$ for $l = 1 \ldots L$ in computation time proportional to L instead of 2^L. This is possible because the L coalescence indicators for individual k each affect different loci and are conditionally independent given T_k and $\mathbf{J}$.

After accepting or rejecting the new value of T_k with $I_{\cdot k}$ integrated out, we must choose new values for $\mathbf{I}_{\cdot k}$ given the chosen value of T_k. Because of their conditional independence, we may separately sample each coalescence indicator I_{lk} for $l = 1 \ldots L$ from its full conditional given the chosen value of T_k. This completes the collapsed MH proposal.

Appendix D

Analysis of Simulated Data

Simulations

Our simulator (https://github.com/skumagai/selfingsim) was developed using simuPOP, publicly available at http://simupop.sourceforge.net/. It explicitly represents $N = 10,000$ individuals, each bearing two genes at each of L unlinked loci. Mutations arise at locus l at scaled rate θ_l (Equation 3), in accordance with the infinite-alleles model.

We assigned to uniparental proportion s^* values ranging from 0.01 to 0.99, with half of the $L = 32$ loci assigned scaled mutation rate $\theta = 0.5$ and the remaining loci assigned $\theta = 1.5$.

We conducted 10^2 independent simulations for each assignment of s^*. Each simulation was initialized with each of the $2N \times 32$ genes representing a unique allele. Most of this maximal heterozygosity was lost very rapidly, with allele number and allele frequency spectrum typically stabilizing well within $10N$ generations. After $20N$ generations, we recorded the realized population, from which 100 independent samples of $L = 32$ loci of size $n = 70$ were extracted. From this collection, we randomly chose $L = 6$ loci and subsampled 100 independent samples of size $n = 6$.

Analysis

We applied our Bayesian method, the F_{IS} method, and RMES to 10^2 independent samples from each of 10^2 independent simulations for each assignment of the uniparental proportion s^*. Our Bayesian method is open source and can be obtained at https://github.com/bredelings/BayesianEstimatorSelfing/. We used the implementation of RMES (David *et al.* 2007) provided at http://www.cefe.cnrs.fr/images/stories/DPTEEvolution/Genetique/fichiers%20Equipe/RMES%202009%282%29.zip.

GENETICS

Supporting Information
www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.179093/-/DC1

A Bayesian Approach to Inferring Rates of Selfing and Locus-Specific Mutation

Benjamin D. Redelings, Seiji Kumagai, Andrey Tatarenkov, Liuyang Wang, Ann K. Sakai, Stephen G. Weller, Theresa M. Culley, John C. Avise, and Marcy K. Uyenoyama

DOI: 10.1534/genetics.115.179093

File S1

Supporting Information

1 Assessment of the methods

1.1 Comparison of the median of the posterior distribution of the uniparental proportion

We address the relative accuracy of estimates of the uniparental proportion s^* produced by our Bayesian method relative to those produced by `RMES` and the F_{IS} method (27) upon application to simulated data (compare Assessment of Accuracy and Coverage using Simulated Data section). While we summarize our posterior distributions of s^* by the *mode* in the main text, we here use the *median*.

We first address application of the methods to simulated data under the large-sample regime ($n = 70$ individuals, $L = 32$ loci). Except in cases in which the true s^* is very close to 0, the error for `RMES` exceeds the error for our method (Figure S1), a trend that is apparent under under both the large- and small-sample regimes. The error for the F_{IS}-based estimate also exceeds the error for our method. It is largest near $s^* = 0$ and vanishes as s^* approaches 1, a pattern distinct from `RMES`.

Both `RMES` and our method show positive bias upon application to data sets for which the true uniparental proportion s^* is close to zero and negative bias for s^* close to unity. This trend reflects that both methods yield estimates of s^* constrained to lie between 0 and 1. In contrast, the F_{IS}-based estimate (27) underestimates s^* throughout the range, even near $s^* = 0$ ($\widehat{F_{IS}}$ is not constrained to be positive). Our method has a bias near 0 that is substantially larger than the bias of `RMES`, and an error that is slightly larger. A major contributor to this trend is that our Bayesian estimate is represented by only the median of the posterior distribution of the uniparental proportion s^*.

Figure S2 indicates that for data sets generated under a true value of s^* of 0 (full random outcrossing), the posterior distribution for s^* has greater mass near 0. We suggest that the bias shown near $s^* = 0$ merely represents uncertainty in the posterior distribution for s^* and not any preference for incorrect values. We note that our method assumes that the data are derived from a population reproducing through a mixture of self-fertilization and random outcrossing. Assessment of a model of complete random mating ($s^* = 0$) against the present model ($s^* > 0$) might be conducted through the Bayes factor.

Figure S3 indicates that all methods show increased error upon application of smaller samples ($n = 10$ individuals, $L = 6$ loci), as expected. Comparison of our method and `RMES` show trends qualitatively similar to the large-sample case: positive bias upon application to data sets for which the true uniparental proportion s^* is close to zero and negative bias for s^* close to unity, with less error exhibited by our method throughout the range of the uniparental proportion (s^*).

1.2 Comparison of the median, mode, and mean

In addition to the mode (Figure 2) and median (Figure S1), error might also be assessed by consideration of the mean of the posterior distribution of s^*. Figure S4 suggests that the bias and root-mean-squared (rms) error of the mode, median, and mean exhibit different properties. For example, the posterior mode shows smaller bias throughout the parameter range, but the median and mean show smaller rms error for s^* near the boundaries (near 0 or 1). That the posterior mode does not display large bias near $s^* = 0$ is consistent with our suggestion that the larger error of the mean in that region reflects higher uncertainty.

1.3 Frequentist coverage

As for the 95% BCIs (Figure 4), Figure S5 indicates that BCIs of different nominal values (0.5, 0.75, 0.9, 0.95, and 0.99) display the same pattern, with coverage exceeding the desired value for intermediate true s^* values and dipping below the desired value for very high values of s^*. Coverage is closer to the nominal value for the 0.99 and 0.95 levels than for the 0.5 level.

2 Data analysis

2.1 Self-fertilizing vertebrate

Figure S6 shows the posterior distributions of number of generations since the most recent outcross event (13) for each sampled individual in the highly inbred BP population of *K. marmoratus*. Figure S7 shows the posterior distributions for the more outbred TC population.

Figures S8 and S9 present posterior distributions of locus-specific mutation rates for the BP and TC populations, respectively. For each locus, Fig. S10 compares the rank order of its median mutation rate estimated from the BP data set versus that from the TC data set. If a relationship exists between the mutation rates estimated from the datasets, it appears to be diffuse.

2.2 Gynodioecious plant

Figure S11 presents posterior distributions for the uniparental proportion (s_G), the proportion of females among reproductives (p_f), the proportion of seeds set by hermaphrodites by self-pollen ($\tilde{s}$), and the viability of uniparental offspring relative to biparental offspring (τ).

Figure S12 presents the inferred number of generations since the most recent outcross event T_k (13) for each individual k.

Figure S13 presents posterior distributions for locus-specific mutation rates inferred from the *S. salicaria* data set. The loci appear to have similar posterior medians.

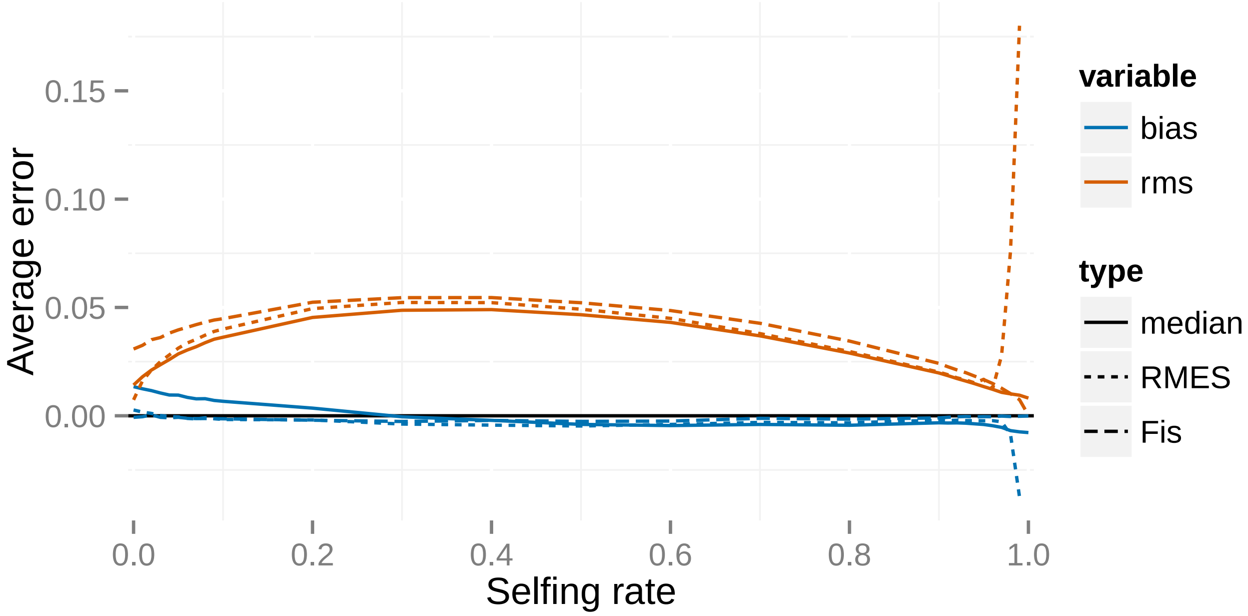


Figure S1 Errors for the full likelihood (posterior median), RMES, and F_{IS}-based (27) methods for a large simulated sample ($n = 70$ individuals, $L = 32$ loci). In the legend, rms indicates the root-mean-squared error and bias the average deviation. Averages are taken across simulated data sets at each true value of s^*.

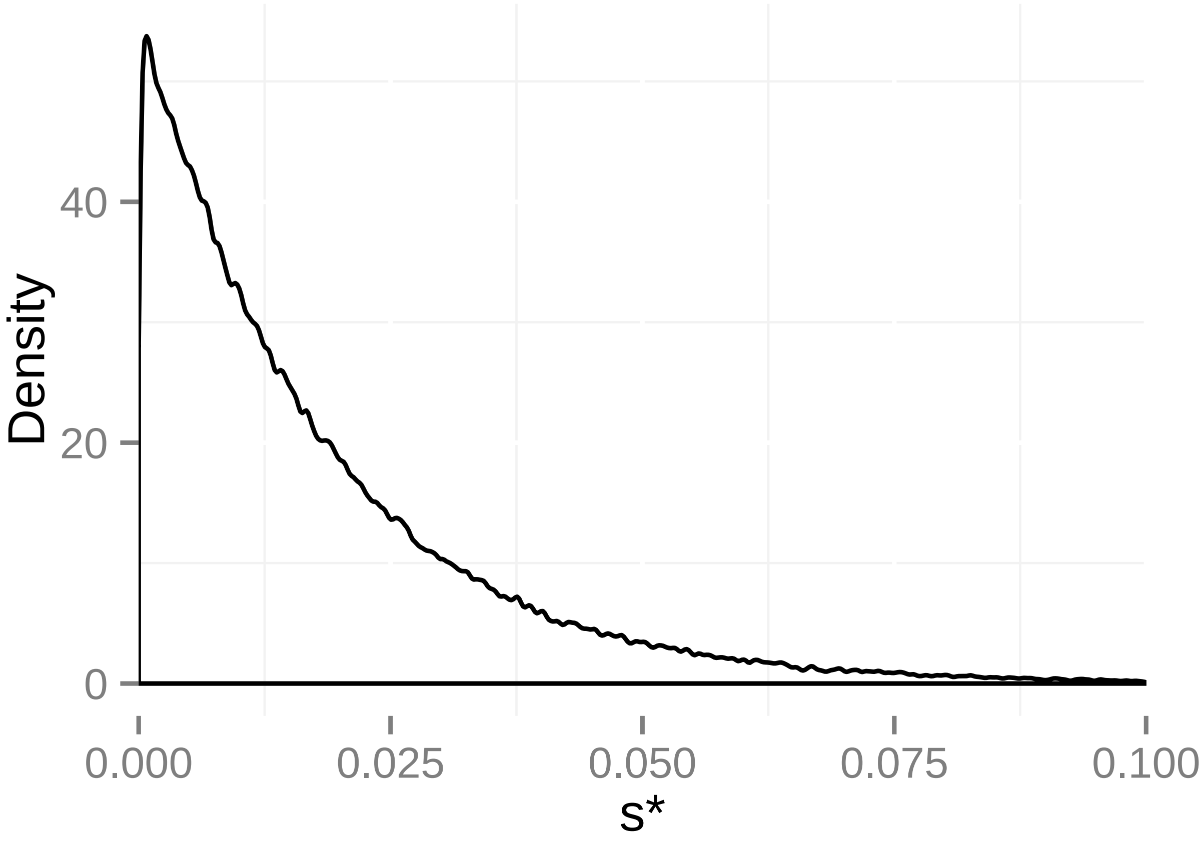


Figure S2 Average posterior density of the uniparental proportion (s^*) inferred from simulated data generated under the large sample regime ($n = 70$, $L = 32$) with a true value of $s^* = 0$. The average was taken across posterior densities for 100 data sets.

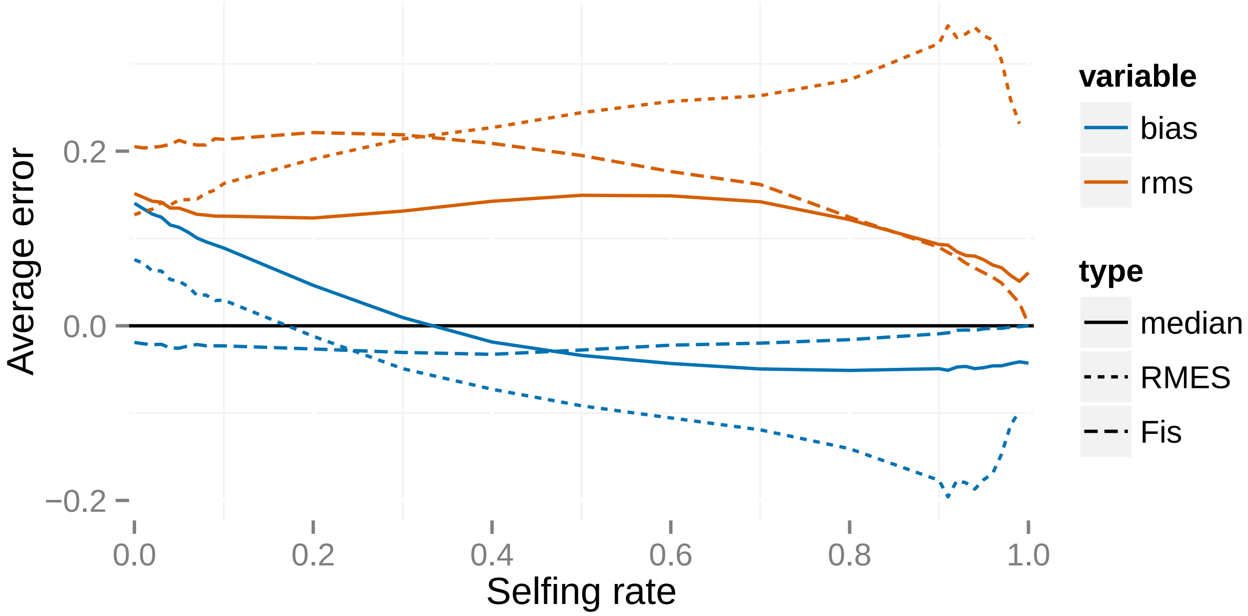
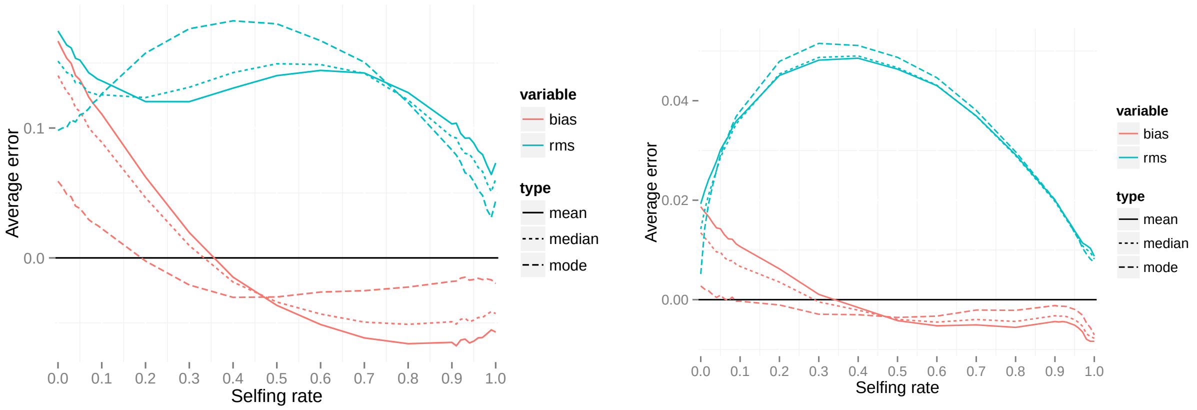


Figure S3 Errors for the full likelihood (posterior median), `RMES`, and F_{IS} methods for a small sample ($n = 10$ individuals, $L = 6$ loci). In the legend, rms indicates the root-mean-squared error and bias the average deviation. Averages are taken across simulated data sets at each true value of s^*.



(a) $n = 10,\ L = 6$

(b) $n = 70,\ L = 32$

Figure S4 Errors for the posterior mean, posterior median, and posterior mode. Blue curves (rms) indicate the root-mean-squared error, and red curves (bias) the average deviation. Averages are taken across simulated data sets at each true value of the selfing rate s^*.

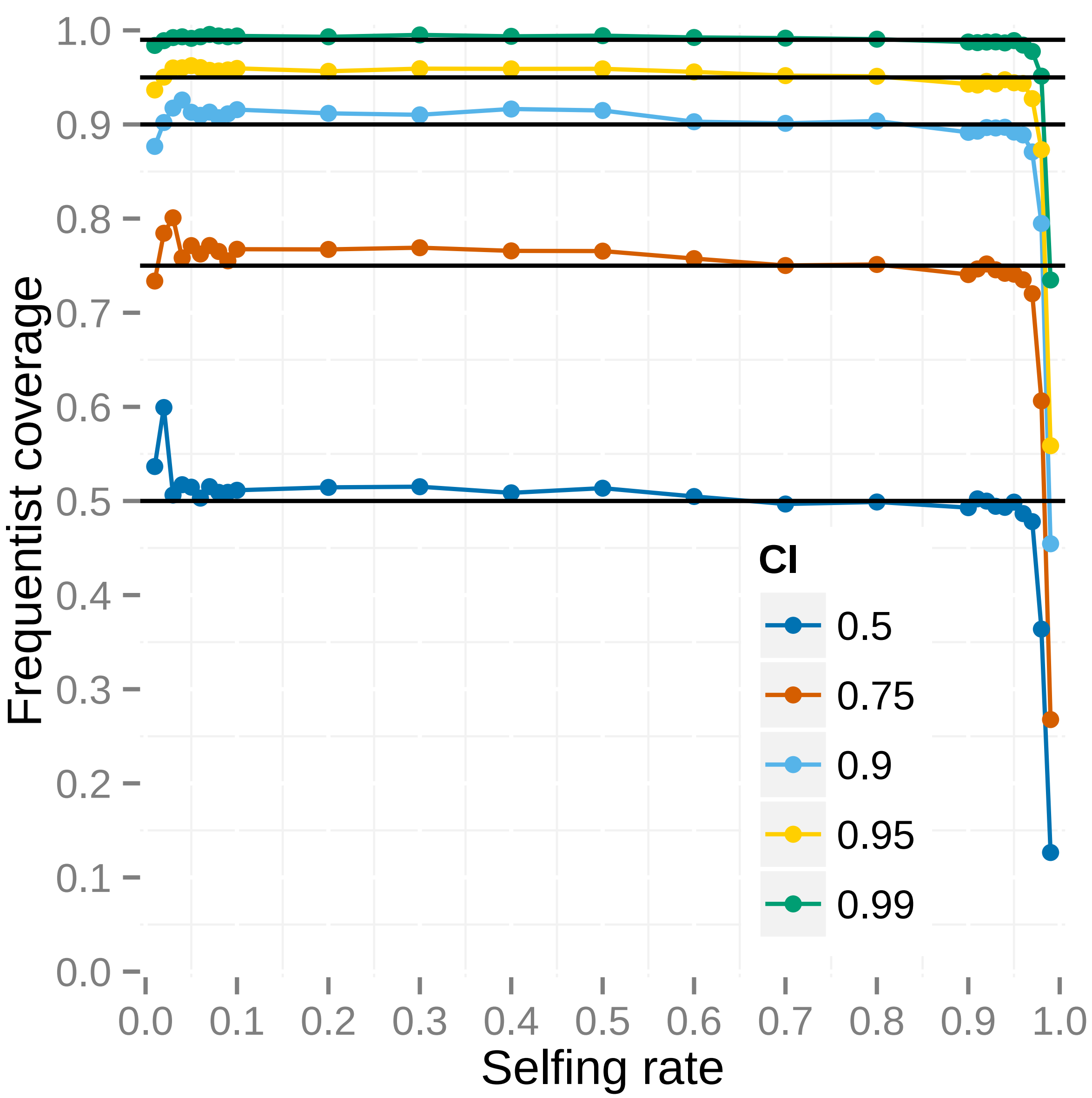


Figure S5 Frequentist coverage for Bayesian credible intervals at different levels of credibility under the large sampling regime $(n = 70, L = 32)$.

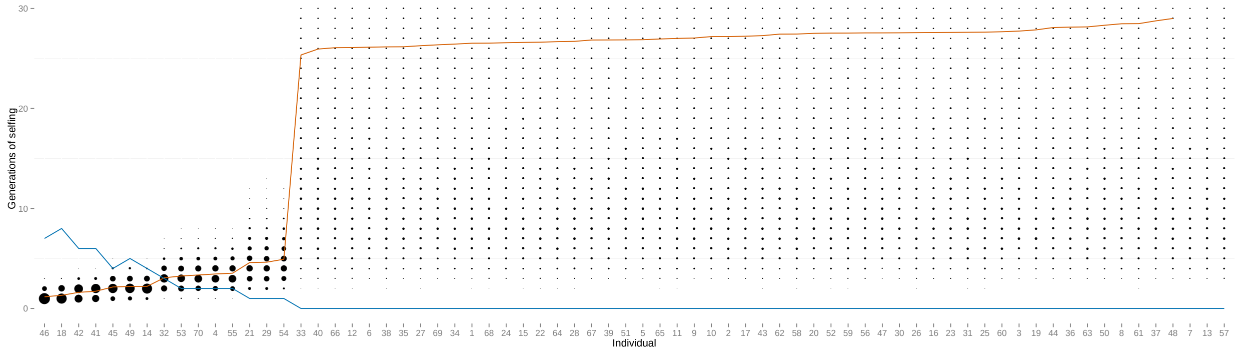


Figure S6 Number of generations since the most recent outcross event in the ancestry of each individual in the sample from the BP population of *K. marmoratus*. The area of each dot indicates the posterior probability that an individual (X-axis) has the indicated number (Y-axis) of consecutive generations of selfing in its immediate ancestry. The red line indicates the posterior mean number of selfing generations and the blue line indicates the number of heterozygous loci across individuals. The Y-axis is truncated to $[0, 30]$.

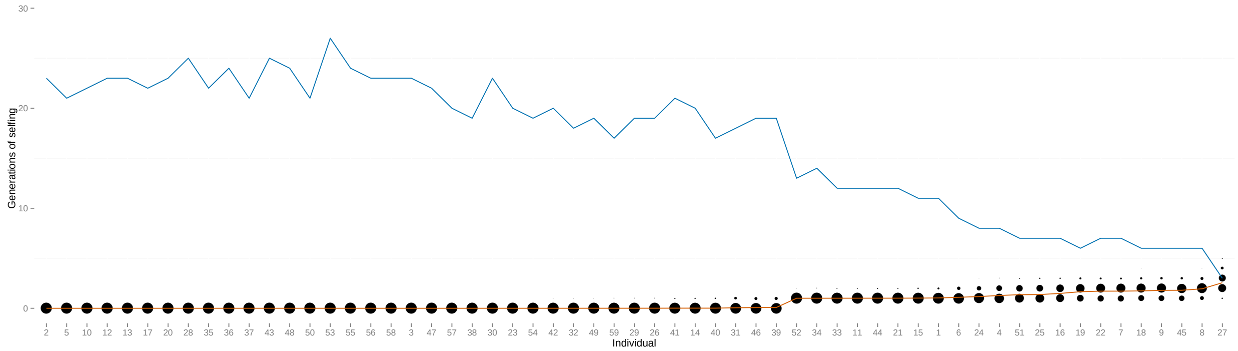


Figure S7 Number of generations since the most recent outcross event in the ancestry of each individual in the sample from the TC population of *K. marmoratus*. Symbols as in Figure S6.

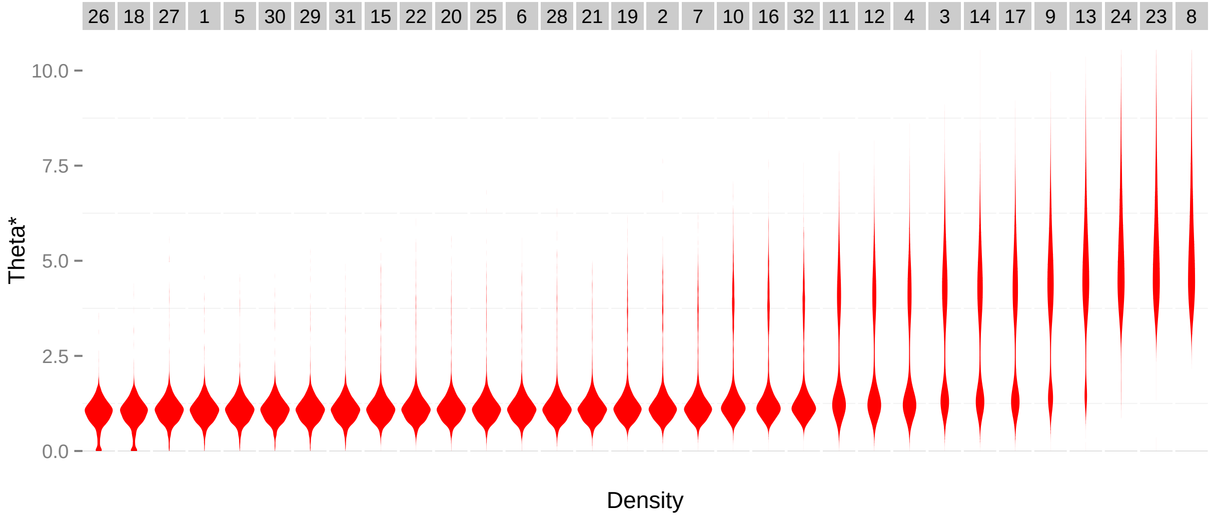


Figure S8 Posterior distributions for mutation rates at each locus in *K. marmoratus* (BP population). For each distribution. the locus name is indicated in the grey shaded box.

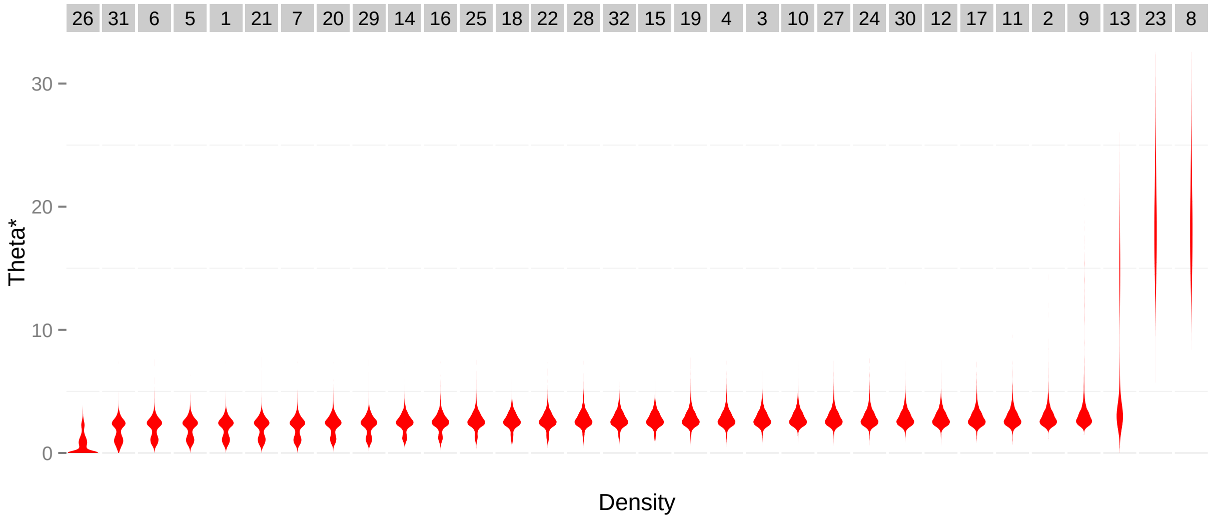


Figure S9 Mutation rates at each locus for *K. marmoratus* (TC population). For each distribution. the locus name is indicated in the grey shaded box.

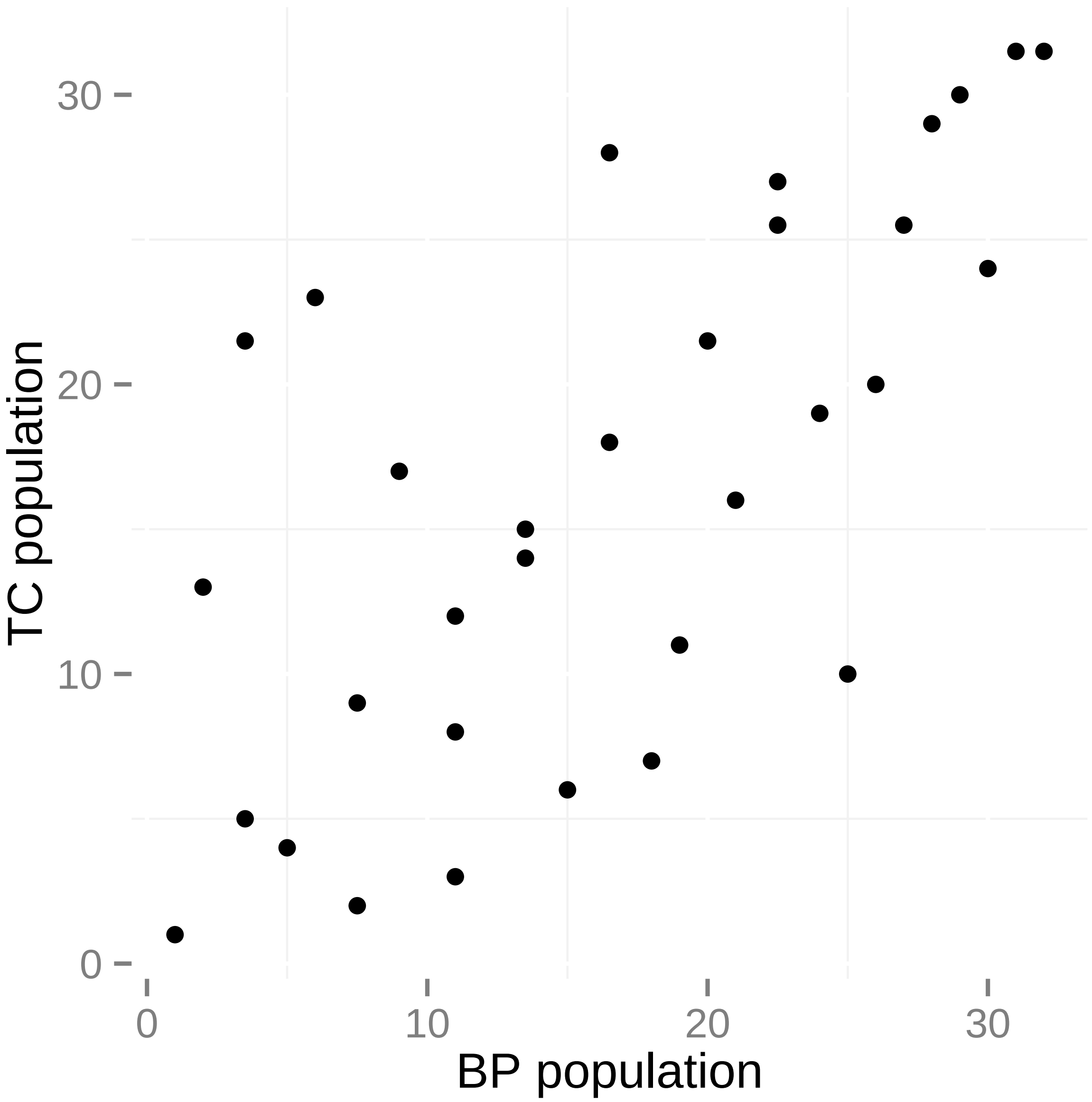


Figure S10 Comparison of rank order of estimated locus-specific mutation rates between the BP and TC populations of *K. marmoratus*. Each dot represents the rank order of the median of the mutation rate of a given locus estimated from the BP data set versus that from the TC data set.

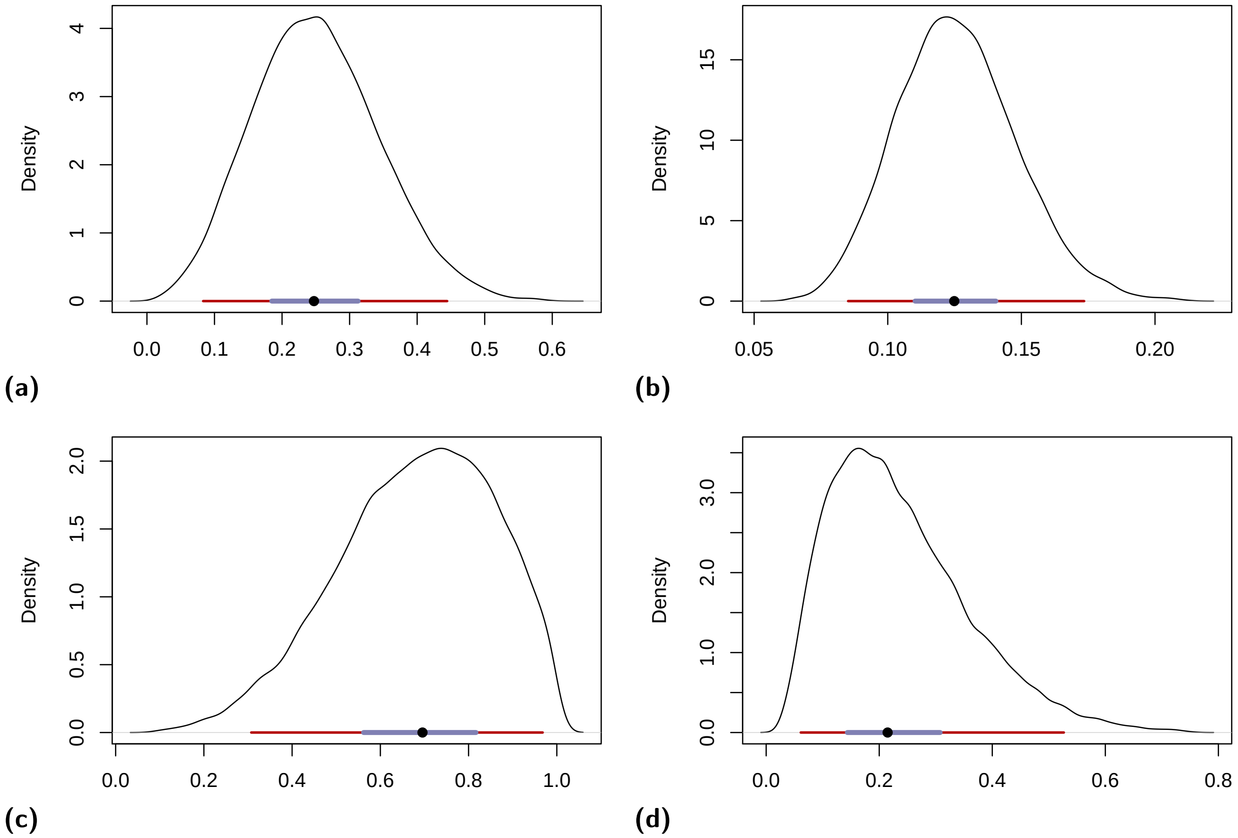


Figure S11 Posterior distributions on (a) s_G, (b) p_f, (c) $\tilde{s}$, and (d) τ for the *Schiedea salicaria* data set. Also shown are 95% BCI (maroon), 50% BCI (slate), and median (black dot).

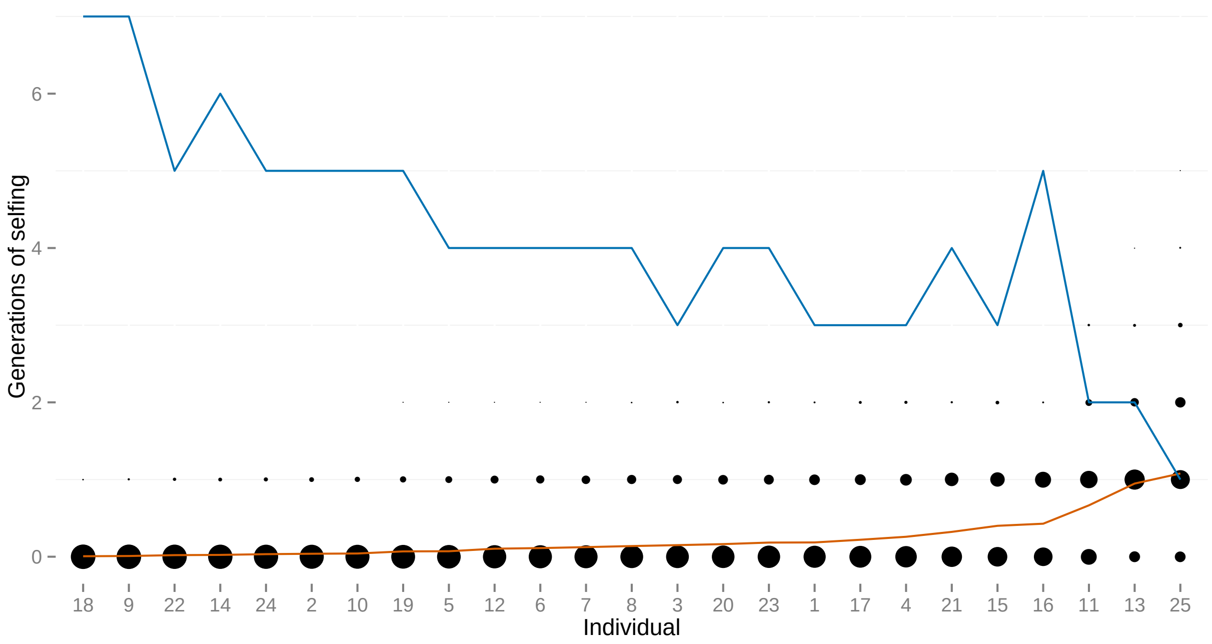


Figure S12 Estimated number of selfing generations for each individual for *S. salicaria*. The area of each dot indicates the posterior probability that a numbered individual (x-axis) has been selfed for a given number of generations (y-axis). For each individual the red line indicates the posterior mean number of selfing generations and the blue line indicates the number of heterozygous loci.

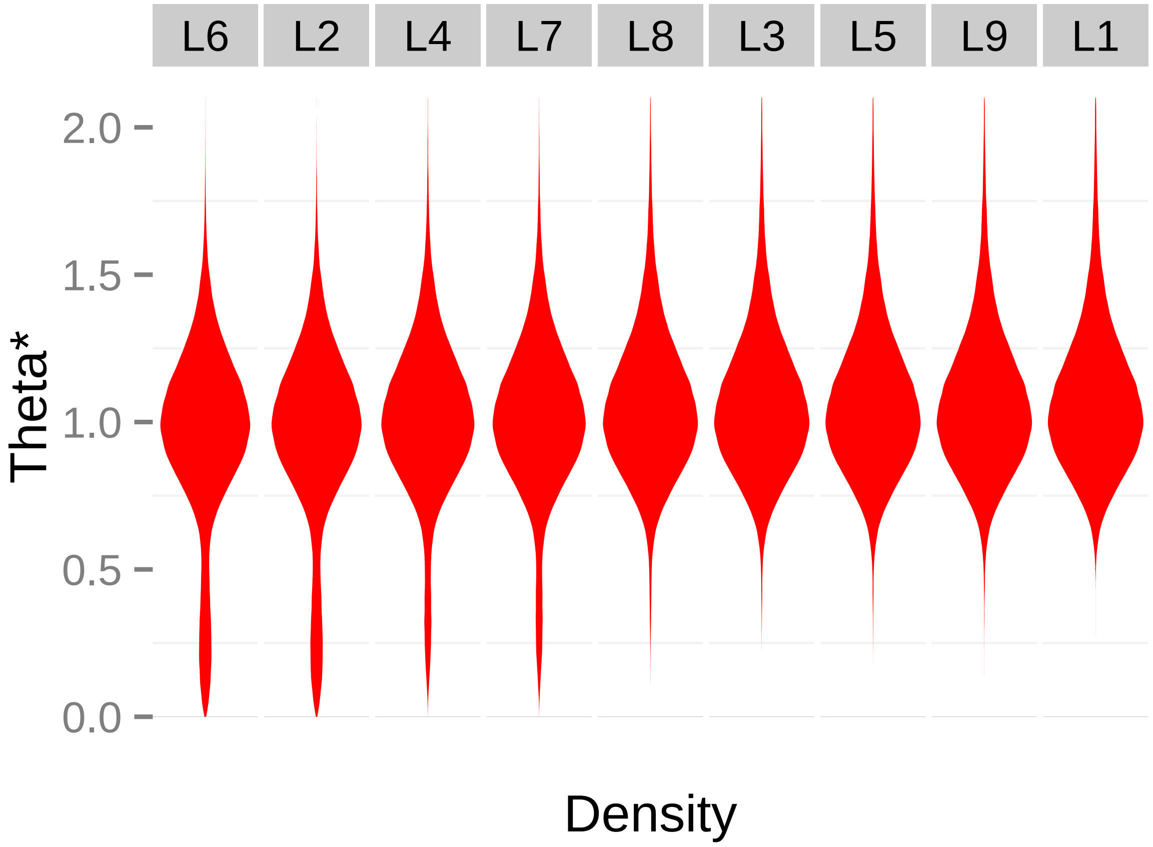


Figure S13 Posterior distributions for mutation rates at locus in *S. salicaria*. For each distribution, the locus name is indicated in the grey shaded box.