Article (Investigation)

Title: Bayesian co-estimation of selfing rate and locus-specific mutation rates for a partially selfing population

Authors:

Benjamin D. Redelings*

Seiji Kumagai*

Liuyang Wang*

Andrey Tatarenkov§

Ann K. Sakai§

Stephen G. Weller§

Theresa M. Culley†

John C. Avise§

Marcy K. Uyenoyama*

*Department of Biology, Box 90338, Duke University, Durham, NC 27708-0338

§Department of Ecology and Evolutionary Biology, University of California, Irvine, Irvine, CA 92697-2525

†Department of Biological Sciences, University of Cincinnati, Cincinnati, OH 45220

Article (Investigation)

Running head: Bayesian estimation of inbreeding

Keywords: selfing rate, Ewens Sampling Formula, Bayesian, MCMC, mating system

Address for correspondence:

Marcy K. Uyenoyama

Department of Biology

Box 90338

Duke University

Durham, NC 27708-0338

USA

Tel: 919-660-7350

Fax: 919-660-7293

e-mail: marcy@duke.edu

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 pure hermaphroditism, androdioecy, and gynodioecy. 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. Among the parameters jointly inferred are the population-wide rate of self-fertilization, locus-specific mutation rates, and the number of generations since the most recent outcrossing event for each sampled individual.

Inbreeding has pervasive consequences throughout the genome, affecting genealogical relationships between genes held at each locus within individuals and among multiple loci. This generation of genome-wide, multilocus disequilibria of various orders transforms 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}), which reflects 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 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), basing the estimation of selfing rates on the distribution of heterozygotes across multiple, unlinked loci, while accommodating 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 progeny for which one or more parents are known, 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 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,

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 directly and jointly with the selfing rate (David *et al.* 2007).

In contrast, our Bayesian method for the analysis of partial self-fertilization derives from a coalescence model that accounts for genetic variation and uses the Ewens Sampling Formula (ESF, Ewens 1972). Our approach replaces the estimation of allele frequencies or heterozygosity (1) by 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 it creates a new mutational class 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 (22a)).

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. Our method infers the number of consecutive generations of self-fertilization in the immediate ancestry of each sampled diploid individual and the probability of coalescence during this period between the lineages at each locus.

In inferring the full likelihood from the observed frequency spectrum of diploid genotypes at multiple unlinked loci, we determine the distributions of the allele frequency spectra ancestral to the sample at the most recent point at which all sampled gene lineages at each locus reside in separate individuals. At this point, the ESF provides the exact likelihood,

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.

Here, we address the estimation of inbreeding rates in populations undergoing pure hermaphroditism, androdioecy (hermaphrodites and males), or gynodioecy (hermaphrodites and females). Our method provides a means for the simultaneous inference of various aspects of the mating system, including the population proportions of sexual forms and levels of inbreeding depression. We apply our method to simulated data sets to demonstrate its accuracy in parameter estimation and in assessing uncertainty. Our application to microsatellite data from the androdioecious killifish *Kryptolebias marmoratus* (Mackiewicz *et al.* 2006; Tatarenkov *et al.* 2012) and to the gynodioecious Hawaiian endemic *Schiedea salicaria* (Wallace *et al.* 2011) illustrates the formation of inferences about a number of biologically significant aspects, including measures of effective population size.

Evolutionary model

We describe our use of the Ewens Sampling Formula (ESF, Ewens 1972) to determine likelihoods based on a sample of diploid multilocus genotypes.

From a reduced sample, formed by subsampling a single gene from each locus from each diploid individual, one could use the ESF 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 (Nordborg and Donnelly 1997; Fu 1997). Observation of diploid genotypes provides information about another parameter: the probability that a random individual is uniparental (uniparental proportion). We describe the dependence of these two composite parameters on the basic parameters of models of pure hermaphroditism, androdioecy, and gynodioecy.

Rates of coalescence and mutation

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

Relative rates of coalescence and mutation: We represent the probability that a random individual is uniparental by s^* and the probability that a pair of genes that reside in distinct individuals descend from the same parent in the immediately preceding generation by $1/N^*$. 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 that individual with equal probability. In the latter case, P is 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. Residence of a pair of lineages in a single individual rapidly resolves either to coalescence, with probability

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

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

$$\frac{(1 + f_c)/2}{N^*} = \frac{1}{N^*(2 - s^*)}. \tag{3}$$

Our model assumes that coalescence and mutation occur on comparable time scales:

$$\begin{aligned} \lim_{\substack{N\to\infty \\ u\to 0}} 4Nu &= \theta \\ \lim_{\substack{N\to\infty \\ N^*\to\infty}} N^*/N &= S, \end{aligned} \tag{4}$$

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, S represents a scaled measure of effective population size (termed "inbreeding effective size" by Crow and Denniston 1988), 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 (Nordborg and Donnelly 1997; Fu 1997). Our model assumes independence between the processes of coalescence and mutation and that these processes occur on a much longer time scale than random outcrossing:

$$1 - s^* \gg u, 1/N^*. \tag{5}$$

For m lineages, each residing in a distinct individual, the probability that the most recent event corresponds to mutation is

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

in which

$$\begin{aligned}\theta^* &= \lim_{\substack{N\to\infty \\ u\to 0}} 2N^*u(2-s^*) = \lim_{\substack{N\to\infty \\ u\to 0}} 4Nu\,\frac{N^*}{N}(1-s^*/2) \\ &= \theta(1-s^*/2)S,\end{aligned} \tag{6}$$

for θ and S defined in (4). In inbred populations, the single parameter of the ESF corresponds to θ^*.

Uniparental proportion and the rate of parent-sharing: In a population comprising N_h hermaphrodites, the rate of parent-sharing corresponds to $1/N_h$, and the uniparental

proportion (s^*) corresponds to

$$s_H = \frac{\tilde{s}\tau}{\tilde{s}\tau + 1 - \tilde{s}}, \tag{7a}$$

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 (4) as N_h, implying

$$S_H \equiv 1. \tag{7b}$$

In androdioecious populations, comprising N_h reproducing hermaphrodites and N_m reproducing males (female-steriles), the uniparental proportion (s^*) is identical to the case of pure hermaphroditism (7)

$$s_A = \frac{\tilde{s}\tau}{\tilde{s}\tau + 1 - \tilde{s}}. \tag{8a}$$

A random gene derives from a male in the preceding generation with probability

$$(1 - s_A)/2,$$

and from a hermaphrodite with the complement probability. A pair of genes sampled from distinct individuals derive from the same parent ($1/N^*$) with probability

$$\frac{1}{N_A} = \frac{[(1 + s_A)/2]^2}{N_h} + \frac{[(1 - s_A)/2]^2}{N_m}. \tag{8b}$$

In the absence of inbreeding ($s_A = 0$), this expression agrees with the classical harmonic mean expression for effective population size (Wright 1969). For the androdioecy model, we assign the arbitrary constant in (4) as the number of reproductives ($N_h + N_m$), implying a scaled rate of coalescence corresponding to

$$\frac{1}{S_A} = \frac{N_h + N_m}{N_A} = \frac{[(1 + s_A)/2]^2}{1 - p_m} + \frac{[(1 - s_A)/2]^2}{p_m}, \tag{8c}$$

for

$$p_m = \frac{N_m}{N_h + N_m} \tag{9}$$

the proportion of males among reproductive individuals. Relative effective number $S_A \in (0, 1]$ takes its maximum for populations in which the effective number N_A, implied by the rate of parent sharing, corresponds to the total number of reproductives ($N_A = N_h + N_m$). At $S_A = 1$, the probability that a random gene derives from a male parent equals the proportion of males among reproductives:

$$(1 - s_A)/2 = p_m.$$

In gynodioecious populations, in which N_h hermaphrodites and N_f females (male-steriles) reproduce, the uniparental proportion (s^*) corresponds to

$$s_G = \frac{\tau N_h a}{\tau N_h a + N_h(1 - a) + N_f \sigma}, \tag{10a}$$

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

$$(1 - s_G)F/2,$$

for

$$F = \frac{N_f \sigma}{N_h(1 - a) + N_f \sigma} \tag{10b}$$

the proportion of biparental offspring that have a female parent. A pair of genes sampled from distinct individuals derive from the same parent ($1/N^*$) with probability

$$\frac{1}{N_G} = \frac{[1 - (1 - s_G)F/2]^2}{N_h} + \frac{[(1 - s_G)F/2]^2}{N_f}. \tag{10c}$$

We assign the arbitrary constant N in (4) as $(N_h + N_f)$, implying a scaled rate of coalescence

of

$$\frac{1}{S_G} = \frac{N_h + N_f}{N_G} = \frac{[1-(1-s_G)F/2]^2}{1-p_f} + \frac{[(1-s_G)F/2]^2}{p_f}, \tag{10d}$$

for

$$p_f = \frac{N_f}{N_h + N_f} \tag{11}$$

the proportion of females among reproductive individuals. As for the androdioecy model, $S_G \in (0,1]$ achieves its maximum only if the proportion of females among reproductives equals the probability that a random gene derives from a female parent:

$$(1-s_G)F/2 = p_f.$$

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

$$\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_L\}, \tag{12}$$

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

$$\mathbf{X}_l = \{\mathbf{X}_{l1,}\mathbf{X}_{l2}, \ldots, \mathbf{X}_{ln}\}, \tag{13}$$

with

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

the genotype at locus l of individual k, with 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

$$\mathbf{T} = \{T_1, T_2, \ldots, T_n\}, \tag{14}$$

for T_k denoting the number of consecutive generations of selfing in the immediate ancestry of the k^{th} individual, and

$$\mathbf{I} = \{I_{lk}\}, \tag{15}$$

for I_{lk} indicating whether the lineages borne by the k^{th} 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 k^{th} individual is geometrically distributed:

$$T_k \sim \text{Geometric}\left(s^*\right), \tag{16}$$

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:

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

Because the pair of lineages at any locus coalesce with probability $\frac{1}{2}$ in each generation of selfing,

$$\Pr(I_{lk} = 0) = \frac{1}{2^{T_k}} = 1 - \Pr(I_{lk} = 1). \tag{17}$$

Figure 1 depicts the recent genealogical history at a locus l in 5 individuals. Individuals 2 and 5 are products of random outcrossing ($T_2 = T_5 = 0$), while the others derive from

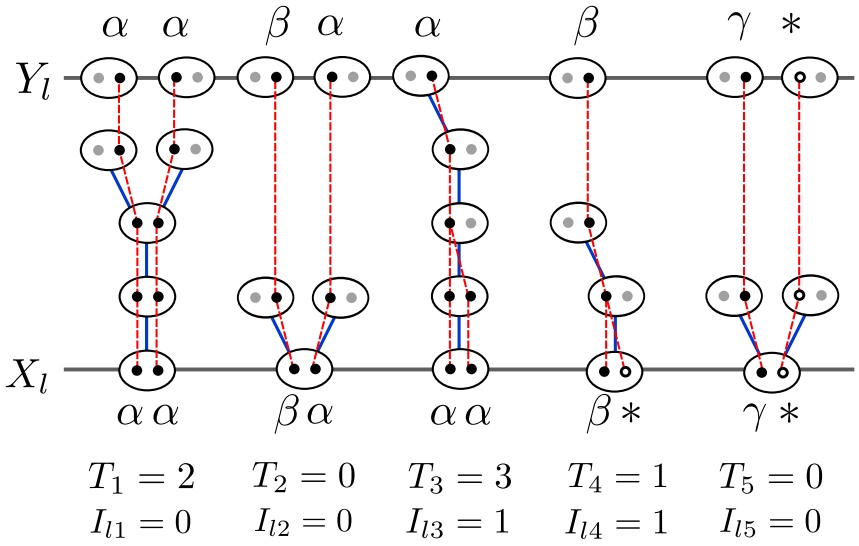

Figure 1 Following the history of the sample ($\mathbf{X}_l$) backwards in time until all ancestors of sampled genes reside in different individuals ($\mathbf{Y}_l$). Ovals represent individuals and dots represent genes. Blue lines indicate the parents of individuals, while red lines represent the ancestry of genes. Filled dots represent sampled genes for which the allelic class is observed (Greek letters) and their ancestral lineages. Open dots represent genes in the sample with unobserved allelic class ($*$). Grey dots 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.

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 are unobserved ($*$), with the unobserved 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. 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

$$\mathbf{Y} = \{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_L\}, \tag{18}$$

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 $\mathbf{X}_l$ (13), 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 ($n \leq m_l \leq 2n$). In Figure 1, for example, $\mathbf{X}_l$ contains 10 genes in 5 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 (5) 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 $\hat{T}$. Accordingly, the coalescence history $\mathbf{I}$ (15) completely determines the correspondence between genetic lineages in $\mathbf{X}$ (12) and $\mathbf{Y}$ (18).

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

$$\Pr(\mathbf{X}|\mathbf{\Theta}^*, s^*) = \sum_{\mathbf{I}} \sum_{\mathbf{T}} \Pr(\mathbf{X}, \mathbf{I}, \mathbf{T}|\mathbf{\Theta}^*, s^*), \tag{19}$$

for

$$\mathbf{\Theta}^* = \{\theta_1^*, \theta_2^*, \ldots, \theta_L^*\} \tag{20}$$

the list of scaled, locus-specific mutation rates, s^* the population-wide uniparental proportion for the reproductive system under consideration (*e.g.*, (7) for the pure hermaphroditism model), and $\mathbf{T}$ (14) and $\mathbf{I}$ (15) 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 (19) 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:

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

Because the times since the most recent outcrossing event $\mathbf{T}$ depend only on the uniparental

proportion s^*, through (16), and not on the rates of mutation $\mathbf{\Theta}^*$,

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

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

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

Further,

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

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 (17).

To compute $\Pr(\mathbf{X}_l|\mathbf{I}_l, \theta_l^*, s^*)$, we incorporate latent variable $\mathbf{Y}_l$ (18), describing the states of lineages at the most recent point at which all occur in distinct individuals (Figure 1):

$$\begin{aligned} \Pr(\mathbf{X}_l|\mathbf{I}_l, \theta_l^*, s^*) &= \sum_{\mathbf{Y}_l} \Pr(\mathbf{X}_l, \mathbf{Y}_l|\mathbf{I}_l, \theta_l^*, s^*) \\ &= \sum_{\mathbf{Y}_l} \Pr(\mathbf{X}_l|\mathbf{Y}_l, \mathbf{I}_l, \theta_l^*, s^*) \Pr(\mathbf{Y}_l|\mathbf{I}_l, \theta_l^*, s^*) \\ &= \sum_{\mathbf{Y}_l} \Pr(\mathbf{X}_l|\mathbf{Y}_l, \mathbf{I}_l) \cdot \Pr(\mathbf{Y}_l|\mathbf{I}_l, \theta_l^*), \end{aligned} \tag{21a}$$

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^* (6).

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 (21a). 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

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

for all $\mathbf{Y}_l$. Any non-zero $\Pr(\mathbf{X}_l|\mathbf{Y}_l,\mathbf{I}_l)$ 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 non-zero $\Pr(\mathbf{X}_l|\mathbf{Y}_l,\mathbf{I}_l)$,

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

Accordingly, expression (21a) reduces to

$$\Pr(\mathbf{X}_l|\mathbf{I}_l,\theta_l^*,s^*)=\sum_{\mathbf{Y}_l:\Pr(\mathbf{X}_l|\mathbf{Y}_l,\mathbf{I}_l)\neq 0}\Pr(\mathbf{Y}_l|\mathbf{I}_l,\theta_l^*), \tag{21b}$$

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 (21b), we use a fundamental property of the ESF (Ewens 1972; Karlin and McGregor 1972): the probability that the next-sampled (i^{th}) gene represents

a novel allele corresponds to

$$\pi_i = \frac{\theta^*}{i - 1 + \theta^*}, \tag{22a}$$

for θ^* defined in (6), and the probability that it represents an additional copy of already-observed allele j is

$$(1 - \pi_i)\frac{i_j}{i - 1}, \tag{22b}$$

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 (22a). Expressions (22) imply that for $\mathbf{Y}_l$ the list of alleles at locus l in order of observance,

$$\Pr(\mathbf{Y}_l|\mathbf{I}_l, \theta_l^*) = \frac{(\theta_l^*)^{K_l}\prod_{j=1}^{K_l}(m_{lj} - 1)!}{\prod_{i=1}^{m_l}(i - 1 + \theta_l^*)}, \tag{23}$$

in which K_l denotes the total number of distinct allelic classes, m_{lj} the number of replicates of the j^{th} allele in the sample, and $m_l = \sum_j m_{lj}$ the number of lineages remaining at time $\hat{T}$ (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 coalescence 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 (23)) provides a sufficient statistic for the estimation of the scaled mutation rate. Because each locus l provides relatively little information about the scaled mutation rate θ_l^* (6), we assume that mutation rates across loci cluster in a finite number of groups. However, we do not know *a priori* the group assignment of loci or even the number of distinct rate classes among the observed loci. We make use of the Dirichlet process prior 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 forms a new group. We assign 0.1 to α of the Dirichlet process, and 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 $\mathbf{\Theta}^*$ (20) 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 are in general non-identifiable 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

$$\begin{aligned}\Pr(\mathbf{X}, \mathbf{\Theta}^*, \mathbf{\Psi}) &= \Pr(\mathbf{X}|\mathbf{\Theta}^*, \mathbf{\Psi}) \cdot \Pr(\mathbf{\Theta}^*) \cdot \Pr(\mathbf{\Psi}) \\ &= \Pr(\mathbf{X}|\mathbf{\Theta}^*, s^*(\mathbf{\Psi})) \cdot \Pr(\mathbf{\Theta}^*) \cdot \Pr(\mathbf{\Psi}),\end{aligned} \tag{24}$$

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 (4) for each locus l, we use (6), incorporating the distributions of $s^*(\mathbf{\Psi})$ and $S(\mathbf{\Psi})$, the scaling factor defined in (4):

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

For the pure hermaphroditism model (7), $\mathbf{\Psi} = \{\tilde{s}, \tau\}$, where $\tilde{s}$ is the proportion of conceptions through selfing, and τ is the relative viability of uniparental offspring. We propose uniform priors for $\tilde{s}$ and τ:

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

For the androdioecy model (8), we propose uniform priors for each basic parameter in $\mathbf{\Psi} =$

$\{\tilde{s}, \tau, p_m\}$:

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

For the gynodioecy model (10), $\mathbf{\Psi} = \{a, \tau, p_f, \sigma\}$, including a the proportion of egg cells produced by hermaphrodites fertilized by selfing, p_f (11) the proportion of females (male-steriles) among reproductives, and σ the fertility of females relative to hermaphrodites. We propose the uniform priors

$$\begin{aligned} a &\sim \text{Uniform}(0,1) \\ \tau &\sim \text{Uniform}(0,1) \\ p_f &\sim \text{Uniform}(0,1) \\ 1/\sigma &\sim \text{Uniform}(0,1). \end{aligned} \tag{27}$$

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 reproducing 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 the Supplementary Online Material.

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

$$\widehat{s^*} = \frac{2\widehat{F_{IS}}}{1+\widehat{F_{IS}}}. \tag{28}$$

In accommodating multiple loci, this estimate incorporates a multilocus estimate for $\widehat{F_{IS}}$ (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 (28) 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-squared 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 median of the posterior distribution of s^*, with the caveat that the mode and mean may show different qualitative behavior (see Supplementary Online Material).

Figure 2 indicates that 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 (28) 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 3

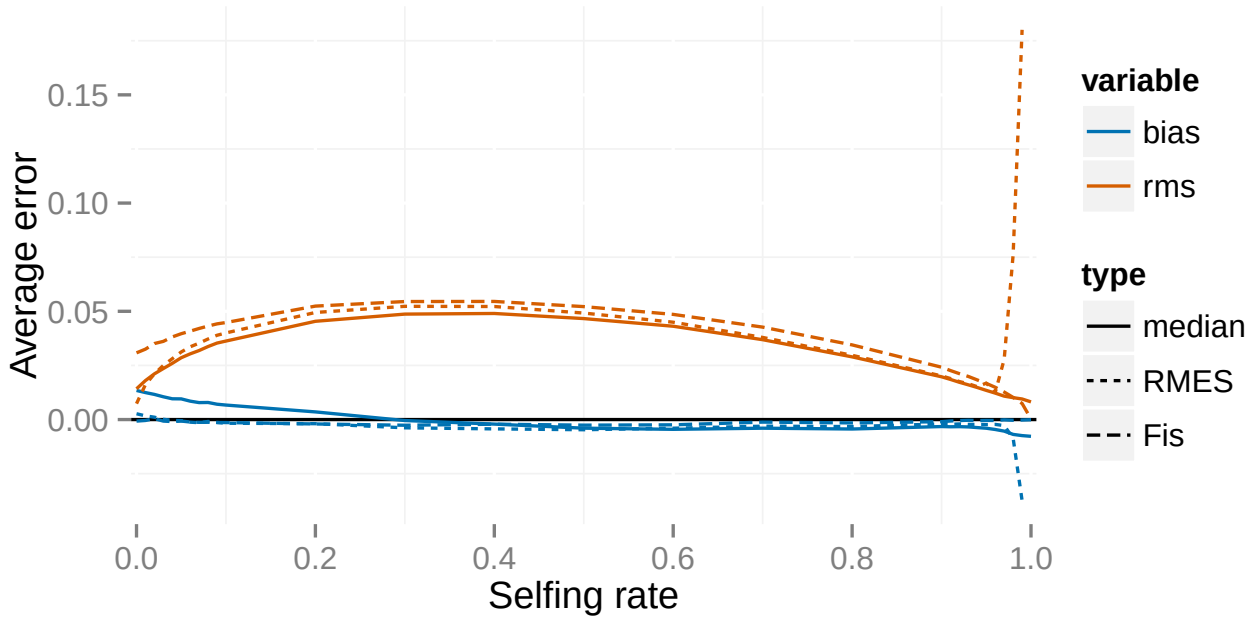

Figure 2 Errors for the full likelihood (posterior median), `RMES`, and F_{IS}-based (28) methods for a large simulated sample ($n = 70$ individuals, $L = 32$ loci). In the legend, rms indicates the root-mean-squared error and bias the average deviation. Averages are taken across simulated data sets at each true value of s^*.

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. Further, as the posterior mode does not display large bias near 0 (Figure S1), we conclude that the bias shown by the median (Figure 2) 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.

Except in cases in which the true s^* is very close to 0, the error for `RMES` exceeds the error for our method under both sampling regimes (Figure 2). `RMES` differs from the other two methods in the steep rise in both bias and rms error for high values of s^*, with the change point occurring at lower values of the uniparental proportion s^* for the small sampling 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

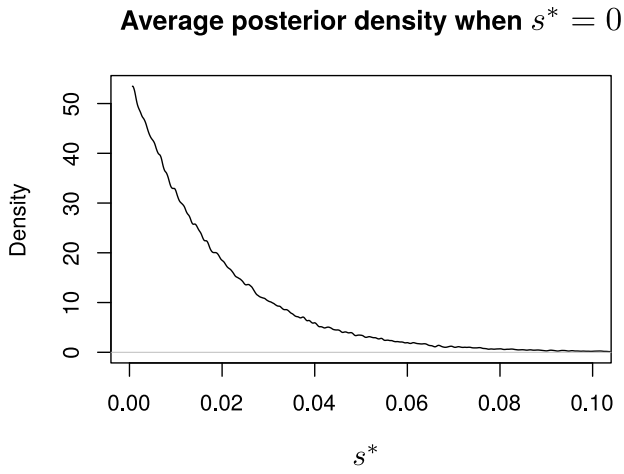

Figure 3 Average posterior density of the uniparental proportion (s^*) inferred from simulated data generated under the large sample regime ($n = 70$, $L = 32$) with a true value of $s^* = 0$. The average was taken across posterior densities for 100 data sets.

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 (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 less than 10% of data sets unanalyzable (fewer than 2 informative loci) even at $s^* = 0.99$ (Figure 4). Under the $n = 10$, $L = 6$ regime, `RMES` discards on average 50% of loci for true s^* values exceeding 0.85, with about 50% of data sets unanalyzable under $s^* \geq 0.94$.

The error for the F_{IS}-based estimate (28) also exceeds the error for our method. It is largest near $s^* = 0$ and vanishes as s^* approaches 1, a pattern distinct from `RMES` (Figure 2).

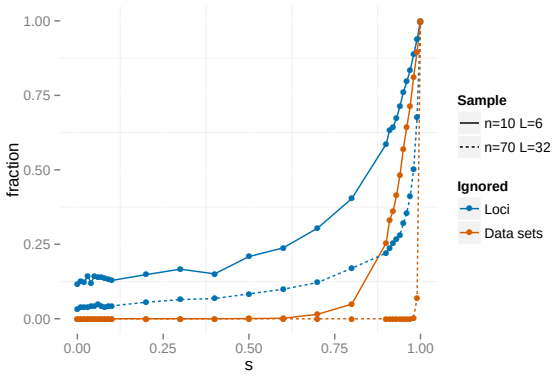

Figure 4 Fraction of loci and data sets that are ignored by `RMES`.

Coverage

We determine the fraction of data sets for which the confidence interval (CI) generated by `RMES` and the Bayesian credible interval (BCI) generated by our method contains 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% CI 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 5 indicates that coverage of the 95% CIs produced by `RMES` are consistently lower than 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 slightly greater than 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 (5) of our underlying model (random outcrossing occurs on a time scale much shorter than the time scales of mutation and coalescence) is likely violated. We observed similar behavior under nominal coverage levels ranging from 0.5 to 0.99 (Supplementary Material).

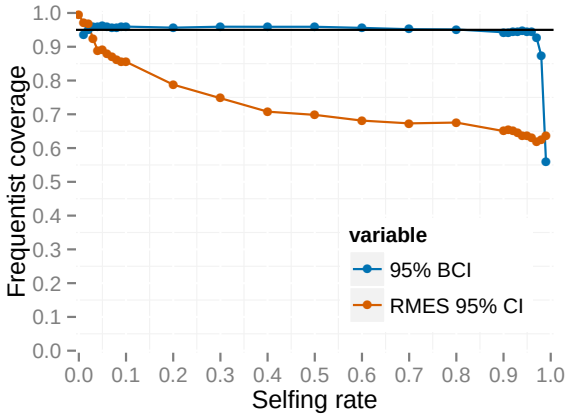

Figure 5 Frequentist coverage at each level of s^* for 95% intervals from `RMES` and the method based on the full likelihood under the large sampling regime ($n = 70, L = 32$). `RMES` intervals are 95% confidence intervals computed via profile likelihood. Full likelihood intervals are 95% highest posterior density Bayesian credible intervals.

Number of consecutive generations of selfing

In order 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

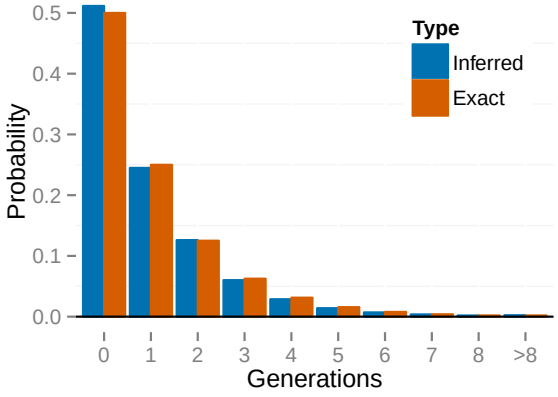

Figure 6 Exact distribution of selfing times under $s^* = 0.5$ compared to the posterior distribution averaged across individuals and across data sets.

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 6). 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

Androdioecious vertebrate

Our analysis of data from the androdioecious killifish *Kryptolebias 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 simultaneously estimates the proportion of males in the population (p_m) together with rates of locus-specific mutation (θ^*) and the uniparental proportion (s_A). We apply the method to two populations, which show highly divergent rates of inbreeding.

Parameter estimation: Our androdioecy model (25) comprises 3 basic parameters, including the fraction of males among reproductives (p_m) and the relative viability of uniparental offspring (τ). Our analysis incorporates the observation of n_m males among n_{total} zygotes directly into the likelihood expression:

$$\Pr(\mathbf{X}, \mathbf{I}, \mathbf{T}, n_m | s^*, \mathbf{\Theta}^*, p_m, n_{total}) = \Pr(\mathbf{X}, \mathbf{I}, \mathbf{T} | s^*, \mathbf{\Theta}^*) \cdot \Pr(n_m | p_m, n_{total}),$$

in which

$$n_m \sim \text{Binomial}(n_{total}, p_m), \tag{29}$$

reflecting that s^* and $\mathbf{\Theta}^*$ are sufficient to account for $\mathbf{X}$, $\mathbf{I}$, and $\mathbf{T}$, and also 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 $\tau = 1$ to permit estimation of the uniparental proportion

s_A 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 on the Florida Keys.

Tatarenkov *et al.* (2012) report 21 males among the 201 individuals collected from various locations in the Florida Keys during this period, consistent with other estimates of about 1% (*e.g.*, Turner *et al.* 1992). Based on the long-term experience of the Tatarenkov–Avise laboratory with this species, we assumed observation of $n_m = 20$ males out of $n_{total} = 2000$ individuals in (29). We estimate that the fraction of males in the population (p_m) has a posterior median of 0.01 with a 95% Bayesian Credible Interval (BCI) of $(0.0062, 0.015)$.

Our estimates of mutation rates (θ^*) indicate substantial variation among loci, with the median ranging over an order of magnitude (ca. 0.5–5.0) (Figure S4, Supplementary Material). 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 7 shows the posterior distribution of uniparental proportion s_A, with a median of 0.95 and a 95% BCI of $(0.93, 0.97)$. This estimate is somewhat lower than F_{IS}-based

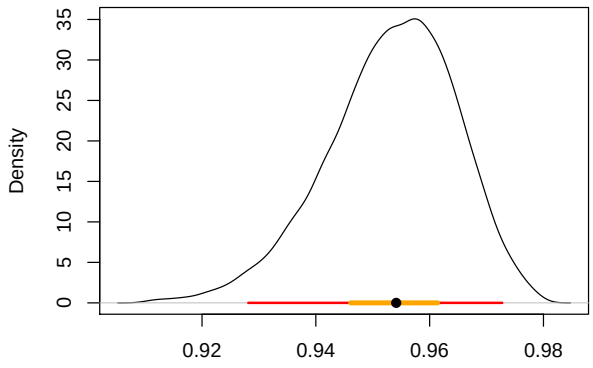

Figure 7 Posterior distribution of the uniparental proportion s_A for the BP population. The median is indicated by a black dot, with a red bar for the 95% BCI and an orange bar for the 50% BCI.

estimate (28) of 0.97, and slightly higher than the `RMES` estimate of 0.94, which has a 95%

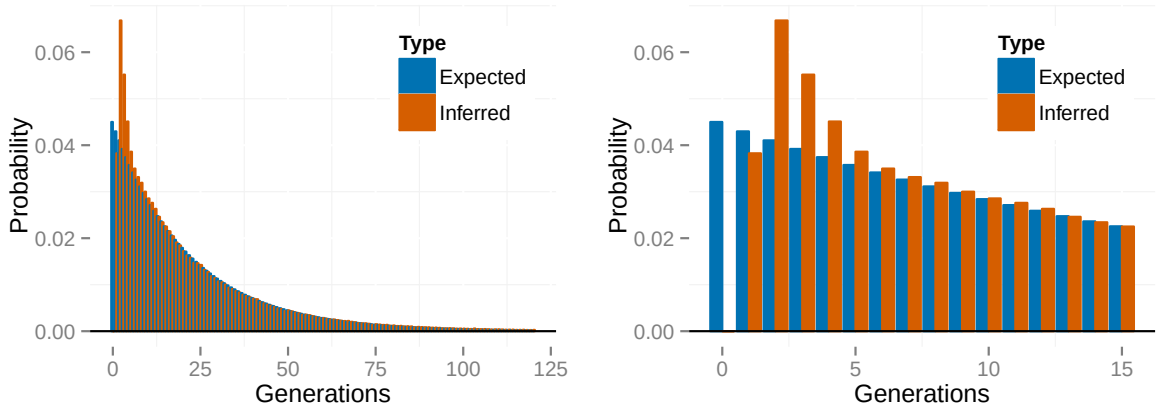

Figure 8 Empirical distribution of number of generations since the most recent outcross event (T) across individuals for the *K. marmoratus* (BP population), averaged across posterior samples. The right panel is constructed by zooming in on the panel on the left. "Expected" probabilities represent the proportion of individuals with the indicated number of selfing generations expected under the estimated uniparental proportion s_A. "Inferred" probabilities represent proportions inferred across individuals in the sample. The first inferred bar with positive probability corresponds to $T = 1$.

Confidence Interval (CI) of $(0.91, 0.96)$. We note that `RMES` discarded from the analysis 9 loci (out of 32) which showed no heterozygosity, even though 7 of the 9 were polymorphic in the sample.

Our method estimates the latent variables $\{T_1, T_2, \ldots, T_n\}$ (14), representing the number of generations since the most recent outcross event in the ancestry of each individual (Figure S5). Figure 8 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 ($s_A \approx 0.95$), this finding suggests that any biparental individuals 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). This data set departs sharply from that of the BP population, showing 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 (29)). Our estimate of the fraction of males in the population (p_m) has a posterior median of 0.17 with a 95% BCI of $(0.11, 0.25)$.

Figure S6 (Supplementary Material) indicates that the posterior medians of the locus-specific mutation rates range over a wide range (ca. 0.5–23). Two loci appear to exhibit a mutation rates substantially higher than other loci, both of which appear to have high rates in the BP population as well (Figure S4).

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_A has a median and 95% BCI of $0.35\,(0.25, 0.45)$ (Figure 9). The

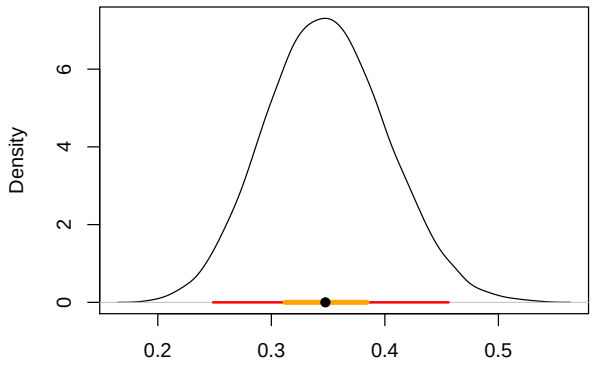

Figure 9 Posterior distribution of the uniparental proportion s_A for the TC population. Also shown are the 95% BCI (red), 50% BCI (orange), and median (black dot).

median again lies between the F_{IS}-based estimate (28) of 0.39 and the `RMES` estimate of 0.33, with its 95% CI of $(0.30, 0.36)$. In this case, `RMES` excluded from the analysis only a single locus, which was monomorphic in the sample.

Figure 10 shows the inferred distribution of the number of generations since the most

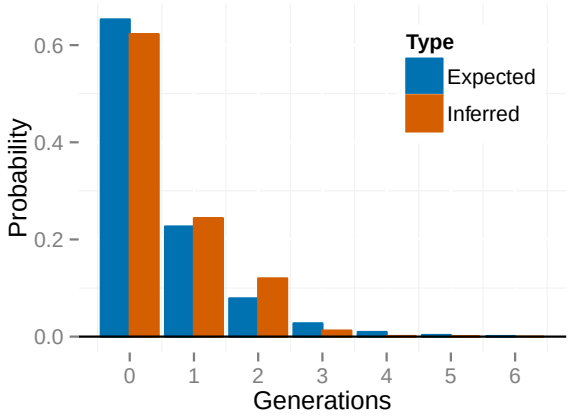

Figure 10 Empirical distribution of selfing times T across individuals, for *K. marmoratus* (Population TC). The histogram is averaged across posterior samples.

recent outcross event (T) across individuals, averaged over posterior uncertainty. In contrast to the BP population, the distribution of selfing time in the TC population appears to conform to the distribution expected under the inferred uniparental proportion (s_A), including a high fraction of biparental individuals ($T_k = 0$). Figure S7 (Supplementary Material) 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 Hawaiiian islands. We analyzed genotypes at 9 microsatellite loci from 25 *S. salicaria* individuals collected from west Maui and identified by Wallace *et al.* (2011) as non-hybrids.

Parameter estimation: Our gynodioecy model (27) comprises 4 basic parameters, including the relative seed set of females (σ) and the relative viability of uniparental offspring (τ). Our analysis of microsatellite data from the gynodioecious Hawaiian endemic *Schiedea salicaria* (Wallace *et al.* 2011) constrained the relative seed set of females to unity ($\sigma \equiv 1$), 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:

$$\tau \sim \text{Beta}(2, 8), \tag{30}$$

the mean of which (0.2) is consistent with the results of greenhouse experiments reported by Sakai *et al.* (1989).

Campbell *et al.* (2010) reported a 12% proportion of females ($n_f = 27$ females among $n_{total} = 221$ individuals). As in the case of androdioecy (29), we model this information by

$$n_f \sim \text{Binomial}(n_{total}, p_f), \tag{31}$$

obtaining estimates from the extended likelihood function corresponding to the product of $\Pr(n_f | n_{total}, p_f)$ and the likelihood of the genetic data. We retain a uniform prior for the proportion of seeds of hermaphrodite set by self-pollen (a).

Results: Figure S10 (Supplementary Material) presents posterior distributions of the basic parameters of the gynodioecy model (10). Our estimate of the uniparental proportion s_G (median 0.247, 95% BCI $(.0791, 0.444)$) is substantially lower than the F_{IS}-based estimate (28) of $s_G = 0.33$. Although `RMES` excluded none of the loci, it gives an estimate of $s_G = 0$, with a 95% CI of $(0, 0.15)$.

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 S8 (Supplementary Material) shows similar posterior medians for across loci.

Figure 11 presents the inferred distribution of the number of generations since the most recent outcross event T across individuals, averaged over posterior uncertainty. In contrast with the analysis of the *K. marmoratus* BP population (Figure 8), the distribution appears to be consistent with the inferred uniparental proportion s_G. Figure S9 (Supplementary Material) presents the posterior distribution of the number of consecutive generations of

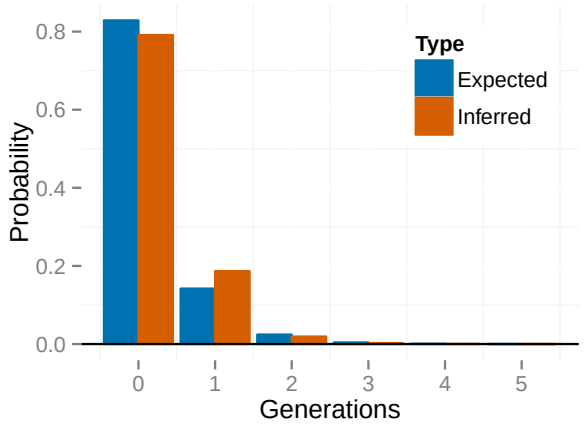

Figure 11 Empirical distribution of selfing times T across individuals, for *S. salicaria*. The histogram is averaged across posterior samples.

selfing in the immediate ancestry of each individual.

Table 1 presents posterior medians and 95% BCIs for the proportion of uniparentals among reproductives (s^*), the proportion of seeds set by hermaphrodites by self-pollen (a), the viability of uniparental offspring 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 ($(1-s_G)F/2$). Comparison of the first (YYY) and fifth (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 first (YYY) and third (YNY) rows indicates that counts of females and hermaphrodites greatly reduce the posterior median of p_f and accordingly change the proportional contribution of females to the gene pool ($(1-s_G)F/2$). The bottom row of the table (NNN), showing a prior estimate for composite parameter s^* of $0.0844\,(0.000797, 0.643)$, illustrates that its induced prior distribution departs from uniform on $(0, 1)$, even though both of its components (a and τ) have uniform priors.

Table 1 Parameter estimates for different amounts of data. Estimates are given by a posterior median and a 95% BCI.

G	F	I	s^*	a	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)

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).

Discussion

We introduce a model-based Bayesian method for the inference of the rate of self-fertilization and other aspects of a mixed mating system. In anticipation of large (even genome-scale) numbers of loci, it uses the Ewens Sampling Formula (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 iconic mating system considered here (pure hermaphroditism, androdioecy, and gynodioecy), permitting insight into various components of the evolutionary process, including effective population size relative to the number of reproductives.

Assessment of the new approach

Accuracy: 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 4). 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 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).

Estimates of the rate of inbreeding produced by our method appear to show greater

accuracy than RMES and the F_{IS}-based method (28) over much of the parameter range (Figure 2). The increased error exhibited under very high rates of inbreeding ($s^* \approx 1$) may reflect violation of our assumption (5) that random outcrossing occurs on a much shorter time scale than mutation and coalescence. Even though our method assumes that the rate of inbreeding lies in $(0, 1)$, the posterior distribution for data generated under random outcrossing ($s^* = 0$) does indicate greater confidence in low rates of inbreeding (Figure 3).

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 (14). This quantity ranges over the non-negative integers, but values assigned to individuals are explored by the MCMC according to their effects on the likelihood.

Frequentist coverage properties: Bayesian approaches afford a direct means of assessing confidence in parameter estimates, and our simulation studies suggest that the Bayesian Credible Intervals (BCIs) generated by our method have relatively good frequentist coverage properties as well (Figure S3). The Confidence Intervals (CIs) reported by the maximum-likelihood method RMES (David *et al.* 2007) appear to perform less well (Figure 5). Although David *et al.* (2007) describe RMES as determining CIs via the profile likelihood method (see Kreutz *et al.* 2013), RMES holds constant parameters other than the uniparental proportion (s^*) instead of reoptimizing them to maximize the likelihood as s^* varies. The result is therefore not a true profile likelihood, which may explain the poor coverage properties of the CIs that RMES provides.

Model fit: Bayesian approaches also afford insight into the suitability of the underlying model. Our method provides estimates of the number of generations since the most recent outcross event in the immediate ancestry of each individual (T). We can pool such estimates of selfing times to obtain an empirical distribution of the number of selfing generations, a procedure particularly useful for samples containing observation of the genotype of many individuals. Under the assumption of a single population-wide rate of self-fertilization, we expect selfing time to have a geometric distribution with parameter 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 10) and from *Schiedea* (Figure 11). In contrast, the empirical distribution for the highly-inbred BP population of *K. marmoratus* (Figure 8) shows an absence of individuals formed by random outcrossing ($T = 0$). That our method accurately estimates T from simulated data (Figure 6) 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. Mis-scoring of heterozygotes as homozygotes due to null alleles or other factors, a possibility directly addressed by `RMES` (David *et al.* 2007), may also in principle contribute to the paucity of outbred individuals.

Components of inference

Locus-specific mutation rates: Our method estimates the scaled mutation rate (4) at each locus using the Dirichlet Process Prior (DPP). This approach improves on existing methods in several ways. First, we estimate a single parameter for each locus instead of estimating multiple allele frequencies per locus as do Enjalbert and David (2000). Second, we estimate for each locus the scaled mutation rate, a fundamental component of the evolution-

ary process, rather than the heterozygosity (1), a random outcome of that process. Third, incorporation of the DPP permits the simultaneous estimation of the number of classes of mutation rates, the mutation rate for each class, and the class membership of each locus. It accords the increased accuracy derived from pooling loci with similar mutation rates without *a priori* knowledge of the partitioning of loci among rate classes or even the number of classes.

Joint inference of mutation and inbreeding rates: For the infinite-alleles model of mutation, the Ewens Sampling Formula (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^* (6), 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 (16), 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^*.

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 $\boldsymbol{\Theta}^*$ (20) across loci. Any model for which the parameter set $\boldsymbol{\Psi}$ (24) comprises more than one parameter is not fully identifiable from the genetic data alone. In the pure hermaphroditism model (7), 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 $\boldsymbol{\Theta}^*$ have the same likelihood.

For each basic parameter in $\boldsymbol{\Psi}$ 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 encompasses an explicit model of the new information. Our androdioecy model (8), for example, comprises 3 parameters, including the frequency of males among reproductives (p_m) as well as $\tilde{s}$ and τ. In our analysis of microsatellite data from the killifish *Kryptolebias 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 (29). In the absence of information regarding inbreeding depression (τ), we assigned $\tau \equiv 1$ to permit estimation of the uniparental proportion (s^*) under a uniform prior distribution. This assignment does not affect the reliability of our estimates (s^*, $\boldsymbol{\Theta}^*$, p_m, S_A, *etc.*); rather, the analysis is agnostic concerning the influence of the relative viability of inbred offspring (τ) and the rate of self-fertilization ($\tilde{s}$) in determining the probability that

a random individual is uniparental (s^*).

Non-identifiable 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 non-identifiable. Even so, informative priors assist in their estimation through Bayesian approaches, which do not require parameters to be identifiable. To explore the data set from *Schiedea salicaria* (Wallace *et al.* 2011), we use our 4-parameter gynodioecy model (10), the basic parameters of which include the proportion of females among reproductives (p_f), the relative seed set of females (σ), the relative viability of uniparental offspring (τ), and the proportion of seeds of hermaphrodites set by self-pollen (a). In a manner similar to the androdioecy study, our analysis uses an extended likelihood function, modeling the number of females as a binomial random variable (31). In addition, we use earlier experimental evidence to justify the assignment of $\sigma \equiv 1$ (Weller and Sakai 2005) and to develop an informative prior for τ ((30): Sakai *et al.* 1989). This procedure permits estimation of 3 basic parameters, including the proportion of seeds of hermaphrodites set by self-pollen (a).

Beyond estimation of the selfing rate

Our MCMC implementation updates the full set of basic parameters, with likelihoods determined from the implied values of composite parameters s^* and $\mathbf{\Theta}^*$. 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 (a) 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 S (4), 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 S 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 (10) corresponds to Wright's 1969 harmonic mean expression for effective population size and S 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 S reflects reductions in effective size due to inbreeding in addition to differences in numbers of the sexual forms. Figure 12 presents posterior distributions of S for the 3 data sets explored here. These results suggest that relative effective number S 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.

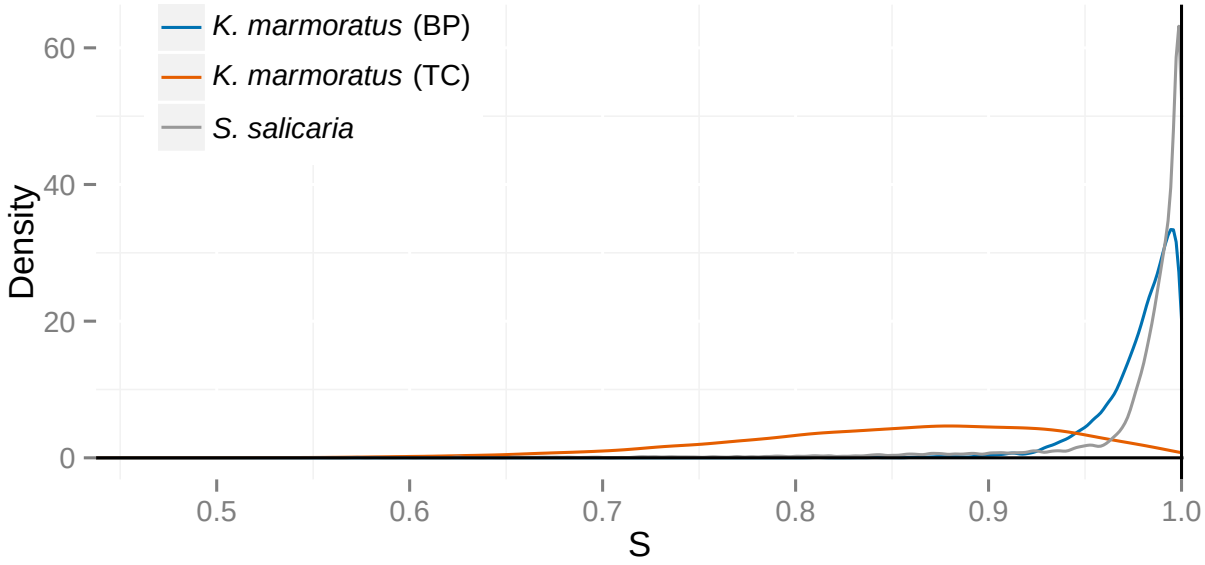

Figure 12 Posterior distributions of relative effective number S (4) for data sets derived from *Kryptolebias marmoratus* (BP and TC populations) and *Schiedea salicaria*.

Acknowledgments

This project was initiated during a sabbatical visit of MKU 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 *Schiedea salicaria*. Public Health Service grant GM 37841 (MKU) provided partial funding for this research.

References

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, and A. K. Dunbar-Wallis, 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. Theoretical population biology **3**: 87–112.

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

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.

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. Theoretical Povulation Biology **68**: 167–177.

Haldane, J., 1924 A mathematical theory of natural and artificial selection. part ii the influence of partial self-fertilisation, inbreeding, assortative mating, and selective fertilisation on the composition of mendelian populations, and on natural selection. Biological Reviews **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. Theoretical Population Biology **3**: 113–116.

Kreutz, C., A. Raue, D. Kaschek, and J. Timmer, 2013 Profile likelihood in systems biology. FEBS Journal **280**: 2564–2571.

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 U S A **103**: 9924–9928.

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. American Journal of Botany pp. 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, T. Roy, W. L. Wagner, and M. Nepokroeff, 2011 Asymmetrical gene flow in a hybrid zone of Hawai-

ian Schiedea (Caryophyllaceae) species with contrasting mating systems. PLoS ONE **6**: e24845, doi:10.1371/journal.pone.0024845.

Wang, J., Y. A. EL-KASSABY, and K. Ritland, 2012 Estimating selfing rates from reconstructed pedigrees using multilocus genotype data. Molecular ecology **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. Journal of Evolutionary Biology **18**: 301–308.

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

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

Appendix A The last-sampled gene

We address the probability that the last-sampled gene in a sample of size i represents a novel allele (22a).

Under the infinite alleles model of mutation, a single mutation in a lineage suffices to distinguish a new allele. We denote the last-sampled gene in a sample of size i as the focal 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

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

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

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

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 (28).

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

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

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

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

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

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

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 $F_{IS} \cdot \sum_l (1 - \sum_u p_{lu}^2)$ 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:

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

Our F_{IS}-based estimates (28) incorporate (B.1) and not (B.2).

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 $\mathbf{T}$ (14), coalescence events across individuals and loci since that event $\mathbf{I}$ (15), and model-specific parameters $\mathbf{\Psi}$ (24). The state space also comprises the scaled mutation rates $\mathbf{\Theta}^*$ (20), 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 $i = 1 \ldots L + 4$. In particular, the scaled mutation rate at locus l corresponds to

$$\theta_l^* = Z_{C_l}. \tag{C.1}$$

At any given point in the MCMC, the state of the Markov chain corresponds to $(\mathbf{I}, \mathbf{T}, \mathbf{\Psi}, \mathbf{C}, \mathbf{Z})$.

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. 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 were needed to achieve smooth density plots for the actual data sets.

Transition kernels: Updating of the continuous variables of mutation rates $\{Z_l\}$ (C.1) and model-specific parameters $\mathbf{\Psi}$ (24) uses both Metropolis-Hastings (MH) transition kernels and auto-tuned 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 Metropolis-Hastings (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 rep-

resents $N = 10,000$ individuals, each bearing two genes at each of L unlinked loci. Mutations arise at locus l at scaled rate θ_l (4), in accordance with the 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 $\theta = 1.5$.

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: To 10^2 independent samples from each of 10^2 independent simulations for each assignment of the uniparental proportion s^*, we applied our Bayesian method, the F_{IS} method, and `RMES`. 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.

Table 1 Parameter estimates for different amounts of data. Estimates are given by a posterior median and a 95% BCI.

G	F	I	s^*	a	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)

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).

Figure captions

Figure 1. Following the history of the sample ($\mathbf{X}_l$) backwards in time until all ancestors of sampled genes reside in different individuals ($\mathbf{Y}_l$). Ovals represent individuals and dots represent genes. Blue lines indicate the parents of individuals, while red lines represent the ancestry of genes. Filled dots represent sampled genes for which the allelic class is observed (Greek letters) and their ancestral lineages. Open dots represent genes in the sample with unobserved allelic class ($*$). Grey dots 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.

Figure 2. Errors for the full likelihood (posterior median), `RMES`, and F_{IS}-based (28) 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 3. 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 4. Fraction of loci and data sets that are ignored by `RMES`.

Figure 5. 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.

Figure 6. Exact distribution of selfing times under $s^* = 0.5$ compared to the posterior distribution averaged across individuals and across data sets.

Figure 7. Posterior distribution of the uniparental proportion s_A for the BP population. The median is indicated by a black dot, with a red bar for the 95% BCI and an orange bar for the 50% BCI.

Figure 8. Empirical distribution of number of generations since the most recent outcross event (T) across individuals for the *K. marmoratus* (BP population), 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 estimated uniparental proportion s_A. "Inferred" probabilities represent proportions inferred across individuals in the sample. The first inferred bar with positive probability corresponds to $T = 1$.

Figure 9. Posterior distribution of the uniparental proportion s_A for the TC population. Also shown are the 95% BCI (red), 50% BCI (orange), and median (black dot).

Figure 10. Empirical distribution of selfing times T across individuals, for *K. marmoratus* (Population TC). The histogram is averaged across posterior samples.

Figure 11. Empirical distribution of selfing times T across individuals, for *S. salicaria*. The histogram is averaged across posterior samples.

Figure 12. Posterior distributions of relative effective number S (4) for data sets derived from *Kryptolebias marmoratus* (BP and TC populations) and *Schiedea salicaria*.

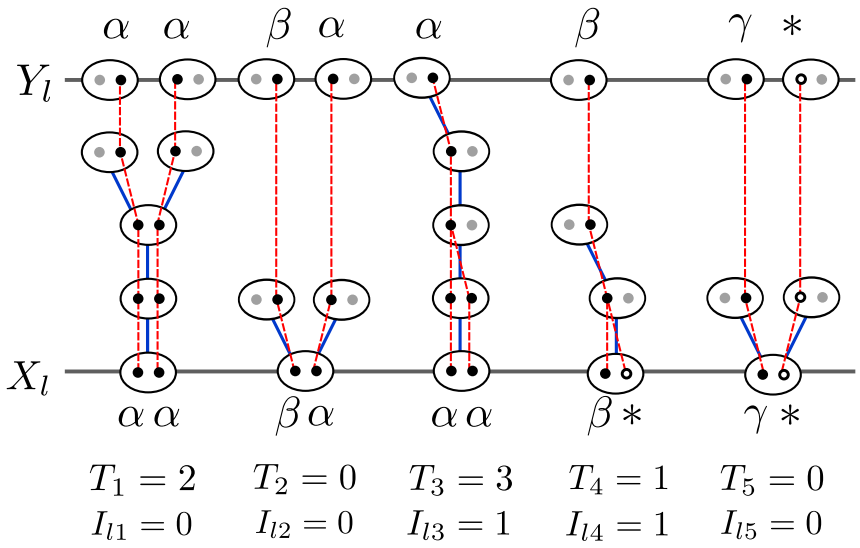

Figure 1 Following the history of the sample ($\mathbf{X}_l$) backwards in time until all ancestors of sampled genes reside in different individuals ($\mathbf{Y}_l$). Ovals represent individuals and dots represent genes. Blue lines indicate the parents of individuals, while red lines represent the ancestry of genes. Filled dots represent sampled genes for which the allelic class is observed (Greek letters) and their ancestral lineages. Open dots represent genes in the sample with unobserved allelic class ($*$). Grey dots 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.

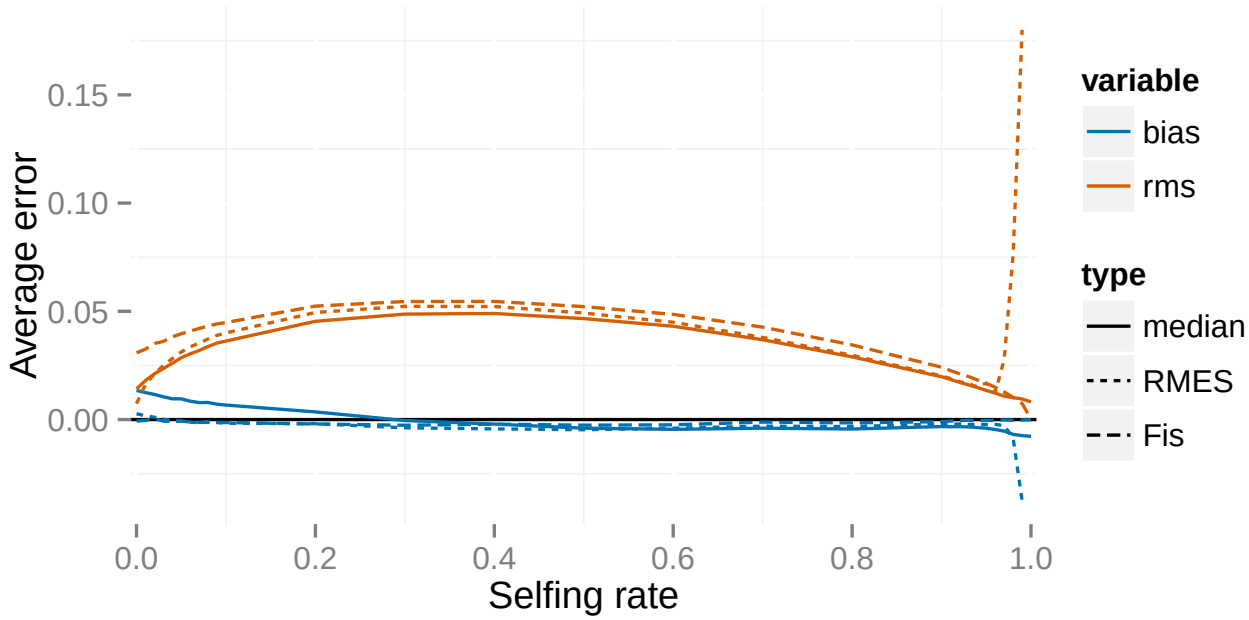

Figure 2 Errors for the full likelihood (posterior median), `RMES`, and F_{IS}-based (28) methods for a large simulated sample ($n = 70$ individuals, $L = 32$ loci). In the legend, rms indicates the root-mean-squared error and bias the average deviation. Averages are taken across simulated data sets at each true value of s^*.

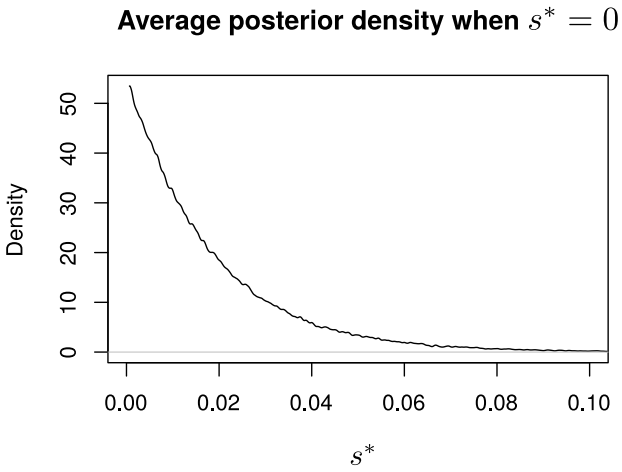

Figure 3 Average posterior density of the uniparental proportion (s^*) inferred from simulated data generated under the large sample regime ($n = 70$, $L = 32$) with a true value of $s^* = 0$. The average was taken across posterior densities for 100 data sets.

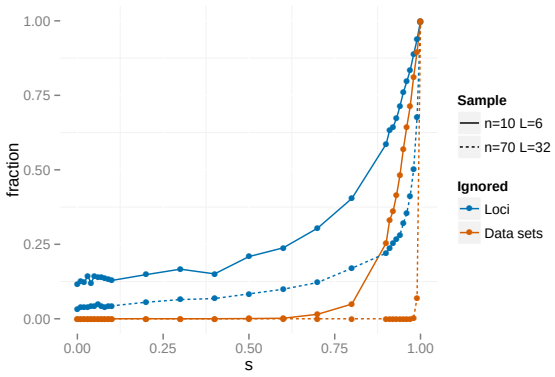

Figure 4 Fraction of loci and data sets that are ignored by `RMES`.

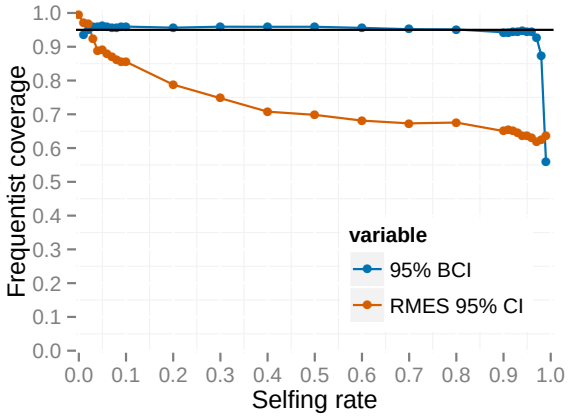

Figure 5 Frequentist coverage at each level of s^* for 95% intervals from `RMES` and the method based on the full likelihood under the large sampling regime ($n = 70, L = 32$). `RMES` intervals are 95% confidence intervals computed via profile likelihood. Full likelihood intervals are 95% highest posterior density Bayesian credible intervals.

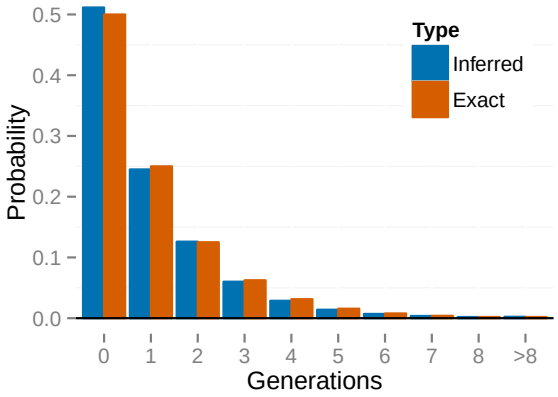

Figure 6 Exact distribution of selfing times under $s^* = 0.5$ compared to the posterior distribution averaged across individuals and across data sets.

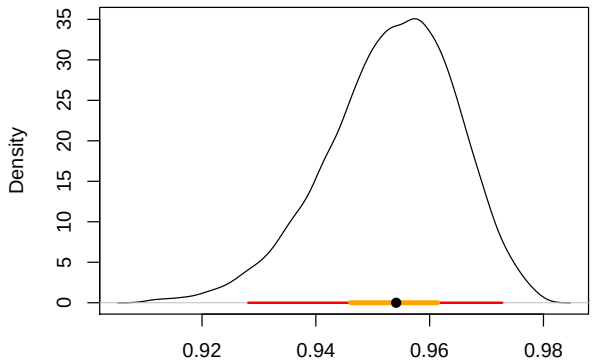

Figure 7 Posterior distribution of the uniparental proportion s_A for the BP population. The median is indicated by a black dot, with a red bar for the 95% BCI and an orange bar for the 50% BCI.

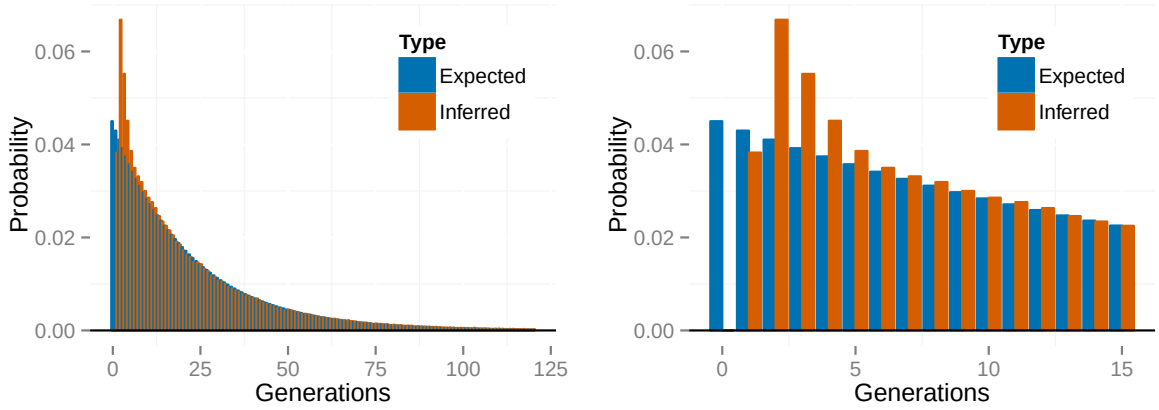

Figure 8 Empirical distribution of number of generations since the most recent outcross event (T) across individuals for the *K. marmoratus* (BP population), averaged across posterior samples. The right panel is constructed by zooming in on the panel on the left. "Expected" probabilities represent the proportion of individuals with the indicated number of selfing generations expected under the estimated uniparental proportion s_A. "Inferred" probabilities represent proportions inferred across individuals in the sample. The first inferred bar with positive probability corresponds to $T = 1$.

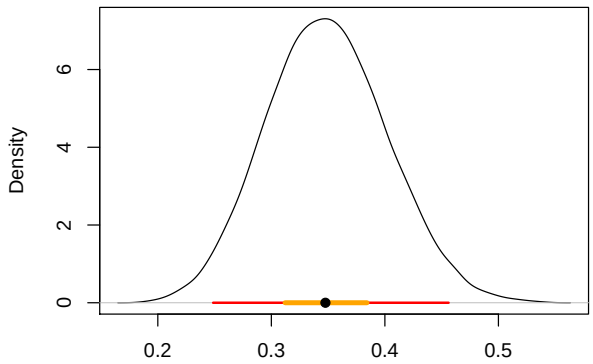

Figure 9 Posterior distribution of the uniparental proportion s_A for the TC population. Also shown are the 95% BCI (red), 50% BCI (orange), and median (black dot).

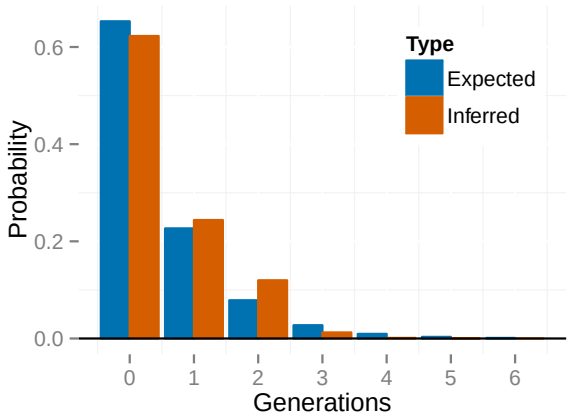

Figure 10 Empirical distribution of selfing times T across individuals, for *K. marmoratus* (Population TC). The histogram is averaged across posterior samples.

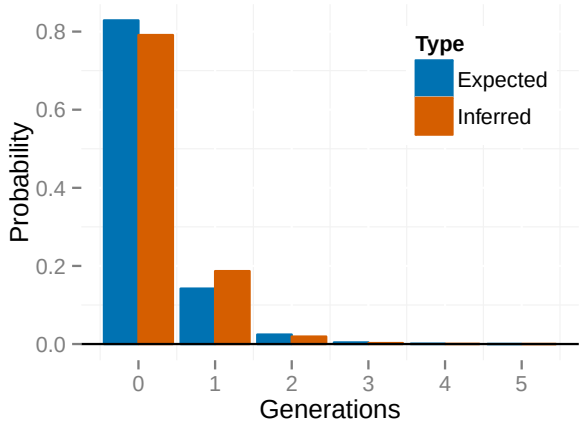

Figure 11 Empirical distribution of selfing times T across individuals, for *S. salicaria*. The histogram is averaged across posterior samples.

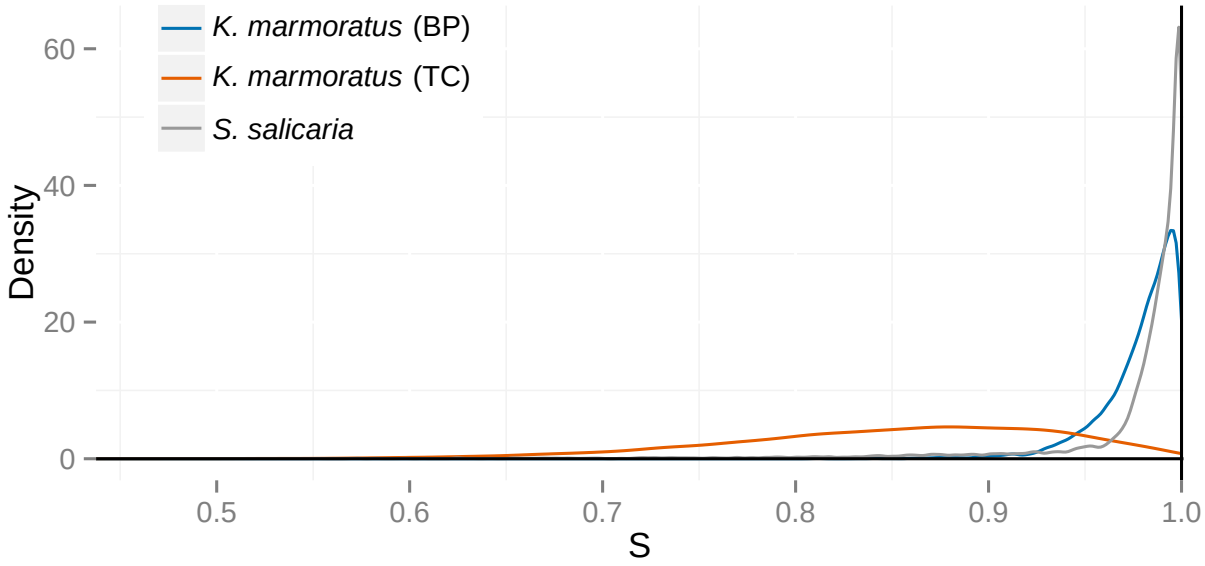

Figure 12 Posterior distributions of relative effective number S (4) for data sets derived from *Kryptolebias marmoratus* (BP and TC populations) and *Schiedea salicaria*.