

SHORT REVIEW

Quantifying inbreeding in natural populations of hermaphroditic organisms

P Jarne and P David

UMR 5175 CEFE, Campus CNRS, Montpellier, France

We review molecular methods for estimating selfing rates and inbreeding in populations. Two main approaches are available: the population structure approach (PSA) and progeny-array approach (PAA). The PSA approach relies on single-generation samples and produces estimates that integrate the inbreeding history over several generations, but is based on strong assumptions (for example, inbreeding equilibrium). The PSA has classically relied on single-locus inbreeding coefficients averaged over loci. Unfortunately PSA estimates are very sensitive to technical problems such as the occurrence of null alleles at one or more of the loci. Consequently inbreeding might be substantially overestimated, especially in outbred populations. However, the robustness of the PSA has recently been greatly improved by the development of multilocus methods free of such bias. The PAA, on the other hand, is based on the comparison between offspring and mother genotypes. As a consequence, PAA

estimates do not reflect long-term inbreeding history but only recent mating events of the maternal individuals studied ('here and now' selfing). In addition to selfing rates, the PAA allows estimating other mating system parameters, including biparental inbreeding and the correlation of selfing among sibs. Although PAA estimates could also be biased by technical problems, incompatibilities between the mother's genotype and her offspring allow the identification and correction of such bias. For all methods, we provide guidelines on the required number of loci and sample sizes. We conclude that the PSA and PAA are equally robust, provided multilocus information is used. Although experimental constraints may make the PAA more demanding, especially in animals, the two methods provide complementary information, and can fruitfully be conducted together. *Heredity* (2008) **100**, 431–439; doi:10.1038/hdy.2008.2; published online 27 February 2008

Keywords: self-fertilization; inbreeding; genetic markers; population structure; progeny-arrays; MLTR

Introduction

Self-fertilization is the fusion of gametes from a single genetic individual. It is common in angiosperms, a group with <10% dioecious species (Renner and Ricklefs, 1995), as well as in other plants, such as mosses (Eppley *et al.*, 2007), and in hermaphroditic animals, such as mollusks or trematodes (Jarne and Auld, 2006). It also occurs in fungi. Estimating selfing rate is an important issue in population and evolutionary biology for two reasons. First, selfing substantially affects the distribution of genetic variation within and among populations, and therefore the response of populations to selection (review in Jarne, 1995; Charlesworth, 2003). Second, the evolution of the selfing rate itself is a topic of central importance (review in Jarne and Charlesworth, 1993; Uyenoyama *et al.*, 1993; Barrett, 2003; Goodwillie *et al.*, 2005). Precise estimates are required to understand the evolution of selfing in relation to reproductive traits (for example, floral traits, copulatory behavior) and inbreeding depression, as well as to test a central prediction of genetic models of mating system evolution, that distributions of selfing rates across species should be

U-shaped (Jarne and Auld, 2006, in animals; Goodwillie *et al.*, 2005, in plants).

The interest in self-fertilization, especially in plants, goes back as far as the mid-nineteenth century (Darwin 1876). However, quantifying the selfing rate became possible only much later. The first attempt might be credited to Jones (1916) who compared parents and offspring genotypes at a dominant morphological marker in tomato, and provided a maximum estimate of the selfing rate. This might be considered as the ancestor of one of the two approaches used today, the progeny-array approach (PAA). The other main approach is based on the analysis of population genetic structure (population structure approach; PSA in what follows). The most well-known PSA method relies on the correspondence between the inbreeding coefficient and the selfing rate (Li, 1955). However, this method became popular only when estimating the inbreeding coefficient was possible, that is with the rise of molecular markers. Curiously, although the selfing rate has been estimated in an extremely large number of studies, these approaches have apparently not been reviewed since the late 1990s (Brown *et al.*, 1989; Brown, 1990).

Our goal here is to fill this gap, that is to review the methods currently used for estimating the selfing rate in populations based on molecular markers, covering progresses in both molecular and statistical methods. After briefly outlining the available markers, we present how they can be used in PSA and PAA. We mention their strengths and weaknesses, including methodological and

Correspondence: Dr P Jarne, UMR 5175 CEFE, Campus CNRS, 1919 route de Mende, 34293 Montpellier cedex 5, France.

E-mail: philippe.jarne@cefe.cnrs.fr

Received 18 July 2007; revised 26 October 2007; accepted 31 December 2007; published online 27 February 2008

technical issues, and argue that they can fruitfully be run together. Note that estimating the selfing rate is often embedded in a wider context, for example in surveys of among-population genetic structure or paternity analysis (see for example, Blouin, 2003), but we restrict our comments to the two approaches mentioned above.

Molecular markers

Mating system parameters, such as the selfing rate, have been estimated using morphological markers with simple genetic basis, especially before the rise of molecular markers. They include flower color in plants (Fryxell, 1957; Cruzan, 1998) or mantle pigmentation in freshwater snails (Vianey-Liaud, 1997). However, selection acts on these markers, often in a rather complex way (for example, Irwin and Strauss, 2005), and it is difficult to get more than one marker per species. The most popular tools are now molecular markers that are assumed to be selectively neutral (see Avise, 2000). The choice of an appropriate marker depends on the same broad criteria whatever the method used (Supplementary Appendix 1): (1) Co-dominant markers are preferable, because it is possible to unambiguously identify heterozygous individuals. Dominant markers can however be used with the PAA (but not with the PSA). (2) Mendelian transmission across generations is required, and is usually assumed for molecular markers (for example, Selkoe and Toonen, 2006). Deviations from co-dominance and Mendelian segregation can in principle be evaluated using progeny-arrays. (3) The precision of all estimates increases with genetic diversity, meaning that highly polymorphic markers are preferred. This requirement is often difficult to meet in highly selfing species, which are usually less polymorphic than their outbred counterparts (Jarne, 1995; Charlesworth, 2003).

The three most commonly used markers are allozymes, microsatellites and AFLP (amplified fragment-length polymorphism), and relevant characteristics are highlighted in Supplementary Appendix 1. Allozymes remain by far the most widely used, totaling 82% of entries in the database used by Goodwillie *et al.* (2005), and 97% of that in Jarne and Auld (2006). Based on literature searches (using ISI Web of Knowledge) and Goodwillie *et al.* (2005), about 15–25 estimates of selfing rate in plants have been published yearly since 1990 using allozymes. Allozymes are co-dominant and five to ten polymorphic loci can easily be scored. However, their generally limited polymorphism in highly inbreeding species is a major incentive for developing microsatellites (for example, Viard *et al.*, 1996). Microsatellites exhibit the same general characteristics of allozymes, but are usually more polymorphic (Jarne and Lagoda, 1996; Estoup and Angers, 1998). Although their use remains limited, with about 6% of the 469 entries in Goodwillie *et al.* (2005), it is increasing with about 10 papers per year since 2001. Although 10 loci can easily be characterized through molecular cloning, this is technically more demanding and time consuming than developing allozymes. AFLPs (Vos *et al.*, 1995), on the other hand, do not require cloning. They are considered as dominant markers with two alleles per locus (present/absent), although the probability that different bands actually represent alleles of the same locus is generally unknown.

These limitations might, to a certain extent, be counter-balanced by the large number of loci scored (several tens to hundreds). Dominance imposes PAA as the only possible method for estimating the selfing rate with AFLPs. Very few studies have been published to date (for example, Thompson and Ritland, 2006).

A major problem with molecular markers is technical bias (see Supplementary Appendix 1). Selfing rate analysis assumes strict correspondence between a scored phenotype (for example, a band on a gel) and an allele. Unfortunately even cross-checked data sets are not error free (Hoffman and Amos, 2005; Pompanon *et al.*, 2005). Most technical biases can be seen as problems of partial dominance, whereby heterozygotes are read as homozygotes for one of the two alleles. This is the case for null alleles that affect both allozymes (for example, David *et al.*, 1997) and microsatellites (see Dakin and Avise, 2004; Selkoe and Toonen, 2006). The same effect is produced whenever two bands in a heterozygote are too close to be visually separated, which can occur in the case of band stuttering at microsatellite loci, or when microsatellite alleles compete for amplification (short allele dominance; Selkoe and Toonen, 2006). In addition, even if heterozygosity is correctly detected, erroneous allele identification affects the expected genotypic frequencies, hence the estimate of selfing rate, whatever the approach. For example if allele A is occasionally read as a new allele B, the observed lack of AB heterozygotes will automatically increase the apparent selfing rate. Other technical problems are related to pattern repeatability and Mendelian transmission across generations (for example, alleles scored in offspring but not in parents). Such problems are probably more acute with AFLPs, but should not be underestimated in the case of microsatellites and especially allozymes (for example, in relation to post-translational modifications of proteins). The influence of technical biases on estimates of the selfing rate is considered in more detail below.

On the whole, microsatellites, arguably as technically robust as allozymes, but usually more polymorphic, are currently the most appropriate markers for estimating inbreeding, and their main limitation is their technical and economic cost.

Population structure approach

The population structure approach is based on samples taken directly in natural populations; all individuals sampled are taken from the same generation. We first describe the classical analysis of such data using single-locus inbreeding coefficients, and then more recent multilocus techniques.

Using single-locus inbreeding coefficients

A simple relationship connects selfing rate, S , to the inbreeding coefficient, F (Fyfe and Bailey, 1951; Li, 1955; Charlesworth, 2003):

$$S = \frac{2F}{1 + F} \quad (1)$$

This equation assumes inbreeding equilibrium in an infinite population of diploid organisms in which selfing is the sole source of inbreeding. Outcrossing occurs at random, and there is no selection, especially no inbreeding depression. A similar formula is available for

tetraploids (Ronfort *et al.*, 1998) and situations in which both gametophytic and sporophytic selfing may occur, such as in mosses (see Eppley *et al.*, 2007). A strong advantage of this approach is that F is almost systematically estimated in analyses of population genetic structure, based on co-dominant markers, providing an ample source of data. For example, estimates of selfing in animals are essentially derived from this approach (Jarne and Auld, 2006).

Various methods and softwares are available for estimating F using genotypes at single loci (Ayres and Balding, 1998; Excoffier, 2001; Rousset, 2001), the most popular estimate being that of Weir and Cockerham (1984). Resampling techniques can be used to assess whether F , and hence the selfing rate, differs from 0 (Weir and Cockerham, 1984; Goudet, 1995). The single-locus sampling variance of \hat{F} is a function of F , n (the number of individuals sampled) and allelic frequencies (Curie-Cohen, 1982; see Supplementary Appendix 2). The variance of \hat{S} follows:

$$\text{Var}(\hat{S}) \approx \frac{4 \text{Var}(\hat{F})}{(1 + \hat{F})^4} \quad (2)$$

$\text{Var}(\hat{S})$ is therefore of the order of $4\text{Var}(\hat{F})$ when F is close to 0, and close to $\text{Var}(\hat{F})/4$ when $F=1$. The essential points to remember are that the error on \hat{S} decreases with n and H_e (gene diversity), and generally with S itself. It can be shown that increasing n is more effective than increasing the number of loci for improving the precision in \hat{S} for the same genotyping effort (Supplementary Appendix 2). However, using several unlinked loci is always a good idea because it limits the impact of locus-specific biases (that is nonneutrality or technical problems). Comparing estimates based on different loci provides a simple evaluation of the overall robustness of the method. Variances and confidence intervals across loci can also be obtained (Excoffier, 2001). For example, Weir and Cockerham (1984) proposed using the jackknife. Finally, when possible, loci with large values of H_e should be typed in priority (Supplementary Appendix 2).

Multilocus approaches

The inbreeding coefficient method can combine information from several loci as a weighted average. This approach certainly increases the precision of F estimates, but genotypic associations among loci are not exploited. Such multilocus information is, however, used by more recent methods (see David *et al.*, 2007). In principle, the simplest way to incorporate this information is to find the selfing rate maximizing the likelihood of all observed multilocus genotypes. In practice however, many variables including allelic frequencies at all loci must be estimated as well. A useful shortcut is to evaluate allelic frequencies first, and then to find S assuming that they are known exactly. This is the approach followed, for example by Enjalbert and David (2000), for estimating S in artificial wheat populations whose founders have been genotyped. Criscione and Blouin (2006) proposed a different approach based on individual inbreeding coefficients estimated from actual population samples. The expected distribution of these coefficients in outcrossed and selfed offspring can be simulated and used to estimate the maximum likelihood function of

S conditional on the observed distribution of inbreeding coefficients. However, the method allows at most a single generation of selfing in individual pedigrees, and is therefore not appropriate for high selfing rates.

David *et al.* (2007) proposed a novel method based on the distribution of multilocus heterozygosity that does not require joint estimation of allelic frequencies. Inbreeding generates not only heterozygote deficiencies, but also identity disequilibria, that is correlations of heterozygosity across pairs of loci (Weir and Cockerham, 1973). In other words, inbred individuals tend to be less heterozygous at all loci than outbred ones. If we consider two loci, the frequency of doubly heterozygous genotypes (and that of doubly homozygous genotypes) is higher than expected under independent assortment (Weir and Cockerham, 1973; David *et al.*, 2007):

$$E(h_1 h_2) = E(h_1)E(h_2)(1 + g_2) \quad (3)$$

where h_1 and h_2 are indicators of heterozygosity at the two loci, $E()$ stands for the expectation and g_2 , the second-order heterozygosity disequilibrium (or identity disequilibrium), gives the relative excess of genotypes heterozygous at two loci. David *et al.* (2007) generalized this idea to several loci and derived a multilocus estimator of S assuming inbreeding and linkage equilibrium. The equation

$$g_2 = \frac{S}{(4 - S)(1 - S)} \quad (4)$$

can be used to derive S from g_2 , and g_2 can be estimated from the distribution of multilocus heterozygosity at a set of loci (David *et al.*, 2007). Note that the only extra conditions required to derive (4), compared to (1), is linkage equilibrium. The sampling variance of \hat{S} is comparable to that obtained using inbreeding coefficients and decreases when the number and genetic diversity of loci increase. David *et al.* (2007) also provide a maximum-likelihood estimate of S , based on the same general idea. This method is implemented in the software RMES (available at <ftp://ftp.cefe.cnrs.fr/>).

Estimates of the selfing rate can also be derived from the long-term effect of selfing on effective recombination and hence linkage disequilibria. For example, Cutter (2006) proposed an approach (detailed in Supplementary Appendix 5) using sequence data, assuming both drift/recombination and inbreeding equilibrium, and requiring some knowledge of recombination rates and effective population size. Given these assumptions, this method provides a long-term estimate of S . The number of assumptions and input parameters restrict the use of such methods and make them unlikely to produce precise estimates, especially when population size is limited, except perhaps for low selfing rates (see Supplementary Appendix 4).

Progeny-array approach

This approach is based on the comparison of mother and offspring genotypes. Ignoring mutation, A_1A_2 offspring with an A_1A_1 mother must be outcrossed and their frequency can provide a minimal estimate of the outcrossing rate $(1-S)$, which of course depends on allelic frequencies. This is the essence of the PAA model for a single locus, set out in Supplementary Appendix 6. This model, or variants thereof, has been used to study

plant mating systems since the premolecular era (Fryxell, 1957). Formal maximum likelihood estimation was initiated by Fyfe and Bailey (1951), and later extended to co-dominant loci (allozymes) by RW Allard and associates (for example, Brown and Allard, 1970; Clegg *et al.*, 1978). A key insight was to incorporate multilocus information (Ritland and Jain, 1981; Shaw *et al.*, 1981), and a good part of subsequent developments of the maximum likelihood method is due to the continuous efforts of Ritland (Ritland, 1986, 1989, 2002; Thompson and Ritland, 2006), and others (for example, Schoen and Clegg, 1986). This was paralleled by a marked increase in the number of studies estimating the selfing rate in natural plant populations (see Schemske and Lande, 1985; Goodwillie *et al.*, 2005). However, the PAA began to be used a few years ago only in animals (see Jarne and Auld, 2006).

The PAA relies on several assumptions, including absence of population substructure and of selection between fertilization and sampling (Table 1; Supplementary Appendix 6). The main difficulty in estimating S is that a nonnegligible proportion (say, β) of outcrossed individuals have genotypes compatible with selfing (Table 2). This proportion can be inferred (using estimated allele frequencies) and used to correct estimates of outcrossing rates, as for example proposed by Shaw *et al.* (1981):

$$1 - \hat{S} = \frac{\hat{n}}{(1 - \hat{\beta})N} \quad (5)$$

with n nonambiguously outcrossed offspring out of N in the array. This simple equation illustrates the advantage

of using multilocus information. Large numbers of loci increase the probability of detecting outcrossing, hence increasing n and decreasing β , reduce the uncertainty on S . Such multilocus estimators rely on the same assumptions than the single-locus ones, plus the absence of linkage disequilibria. Following the same logic, Ritland and Jain (1981) proposed a more general, likelihood-based framework for the PAA, which has subsequently been refined (Ritland, 1986, 1989, 2002) (see Supplementary Appendix 6). The model was embedded in a more general framework and some assumptions were relaxed, allowing for a wider set of biological situations and introducing new parameters:

- The initial model was intended for diploids assessed with co-dominant markers in a mixed-mating model. Extensions were developed for tetraploids with tetrasomic inheritance and no double reduction at meiosis (Murawski *et al.*, 1994; Thompson and Ritland, 2006), automixis and apomixis (Thompson and Ritland, 2006), and dominant markers (Ritland, 1990b; Thompson and Ritland, 2006).
- S can be estimated at different sampling levels, for example individuals and/or families (Ritland and El Kassaby, 1985). This opens the possibility of studying stratified populations, for example the case when two sexual morphs might exhibit different selfing rates, such as in gynodioecious plants or aphyllid snails (Ritland, 2002).
- It was already clear in Shaw *et al.* (1981) that differences between single-locus and multilocus estimates provide an estimate of the amount of biparental inbreeding. However, a more formal analysis was

Table 1 Influence of technical problems, biological processes (population structure and history, selection), marker characteristics and sample size on estimates of the selfing rate (and when relevant, on their variance) when using the single-generation single-locus (F), single-generation multilocus or progeny-array approaches

	F	MLH	PAA	Problem
<i>Technical problems</i>				
Null alleles, band stuttering, fuzzy bands, SAD	+	0	+/- ^a	O
Misscoring	+ ?	?	?	
<i>Population structure and history</i>				
Biparental inbreeding	+	+	+	O
Wahlund effect	+	+	+ (Var(S) +)	O
Departure from inbreeding equilibrium	+/- ^b	+/- ^b	0	S
Departure from gametic equilibrium	?	+/-	+	S
Asexual reproduction	—	?	0 ^c	S ^d
<i>Selection (ID)</i>				
	—	—	—	O
<i>Marker characteristics</i>				
Variability	Var(S)—with variability	Var(S)—with variability	+	S
Number of loci (L)	Var(S)—with L	Use more than two loci	Use five to six loci	
<i>Number of individuals (n)</i>				
	Var(S) in $1/n$	Var(S) in $1/n$	Use 10 offspring per family (15 to 20 if mother unknown)	O

Abbreviations: ID, inbreeding depression; MLH, multilocus heterozygosity; PAA, progeny-array approach; SAD, short-allele dominance. The last column indicates whether selfers (S) or outcrossers (O) are more likely to be affected. An example of Wahlund effect is heterogeneity among pollen pools, and departure from inbreeding equilibrium refers to historical variation in S . Most technical problems are indeed related to partial dominance.

+ and - indicate increase and decrease in the mean selfing rate, 0 and ? no and unknown effect.

^aDepending on how conflicting mother-offspring genotypes are dealt with.

^bDepending on past S .

^cFormally taken into account.

^dBut probably rare in selfers. See Supplementary Appendices 1 and 6 for more details.

Table 2 Genotypes of a mother and three progenies at four diploid loci (A–D) with up to three alleles

Locus	A	B	C	D
Mother	A ₁ A ₁	B ₂ B ₂	C ₁ C ₂	D ₁ D ₃
Offspring 1	A ₁ A ₁	B ₂ B ₂	C ₁ C ₂	<u>D₂D₃</u>
Offspring 2	A ₁ A ₁	B ₂ B ₂	C ₂ C ₂	<u>D₁D₁</u>
Offspring 3	A ₁ A ₁	<u>B₁B₂</u>	C ₁ C ₂	<u>D₂D₃</u>

Underlined offspring genotypes are indicative of outcrossing events, removing uncertainty at other loci when considering all loci at once. For example, offspring 1 genotypes at loci A–C can be derived from either selfing or outcrossing. Locus D genotype allows rejecting selfing assuming no mutation (after Shaw *et al.*, 1981).

proposed through the effective selfing model (Ritland, 1984, 1986, 2002), an effective selfing event occurring either through uniparental selfing, or biparental inbreeding. In more genetical terms, effective selfing accounts for the co-ancestry between male and female contributions in outcrossed offspring. This was further elaborated in connection with the correlated-matings model (see below).

- In the basic model, the mother genotype is compared separately to that of each offspring. Further information can be retrieved from the comparison of offspring genotypes, and this is the cornerstone of the correlated-matings model (Ritland, 1989, 2002). The selfing rate is here estimated together with three other parameters, namely r_s (correlation of selfing), r_p (correlation of outcrossed paternity) and paternal F . The correlation of selfing between two sibs is the covariance of selfing divided by $S(1-S)$. r_s can be interpreted as the fraction of sib pairs that are fully selfed, and it tends toward 1 when sibs are all selfed, or all outcrossed within a family. The correlation of outcrossed paternity is the proportion of full sibs among outcrossed sibs. The correlation r_p is simply related to F and the probability of identity by descent for the two paternally derived gametes in the two progenies (Ritland, 1989). The inverse of r_p can be interpreted as the effective number of fathers. A multilocus version of the correlated-matings model has been proposed (Ritland, 2002). One motivation is that the difference between single and multilocus estimates can be used to calculate biparental inbreeding, yet the uncorrected parameter will be an underestimate by a quantity which is a function of the number of loci (among other parameters), potentially biasing comparisons across studies.

One limitation of the PAA is that it is not always possible to obtain progeny-arrays from natural environments (see Supplementary Appendix 6). This is one of the reasons why it has rarely been used in animals (10 entries only in Jarne and Auld, 2006). On the other hand, its popularity among botanists (about 80% of entries in the database of Goodwillie *et al.*, 2005) might be explained by several outstanding features, some of which are not shared with the PSA. (1) The PAA provides estimates of S that are closer to primary selfing rates because offspring are usually sampled at juvenile stages. (2) Estimates of S can be stratified at lower levels, including families and individuals. However, when the variance in S among families is large, a large number of

families should be studied in order to get a precise mean estimate of S . This seems to be about the case: we haphazardly extracted 30 studies, published from 1993 on, from Goodwillie *et al.* (2005). Out of the 32 species studied, the average number of families per population was 25.9 (s.d. 16.91), not much less than the *ca.* 30 individuals generally studied in PSA. In the 30 PAA studies, 17.1 (8.92) offspring were analyzed on average per family (the number was not correlated with the number of families in the study). (3) The PAA provides instantaneous estimates of S , which can be used to test for selective pressures acting at some point in time, whereas PSA estimates integrate various sources of inbreeding over several generations. (4) The comparison between single and multilocus estimates of S in the PAA allows detection of nonselfing sources of inbreeding (Ritland and Jain, 1981; Ritland, 1986, 2002). Note though that large sample sizes are required to obtain reasonable precision. A further reason of the success of the PAA is that the selfing rate, and correlated-matings parameters, can be estimated using the software MLTR (Ritland, 2002; <http://www.genetics.forestry.ubc.ca/ritland/>). The most recent version has been improved to accept large numbers of alleles per locus, a desirable quality when using microsatellites.

Technical and methodological biases

The main methods presented above (the inbreeding-coefficient approach, the multilocus approach of David *et al.* (2007) and the PAA) are not free of bias, and we distinguish technical biases affecting the chosen markers from methodological biases due to violation of model assumptions.

Technical bias

The technical problems affecting molecular markers (see above) produce biased estimates of the selfing rate, especially using inbreeding coefficients and PAA, as these approaches assume that genotypes are known without errors. Null alleles, short allele dominance and band stuttering generate heterozygote deficiencies, increasing F and apparent selfing (Table 1). Misreading alleles has the same effect, not because the observed heterozygosity decreases, but because the expected heterozygosity usually increases. On the other hand, the genotyping errors mentioned above should on average lead to underestimating S in the PAA. The reason is that a random error is more likely to produce a mother–offspring comparison compatible with outcrossing than with selfing.

The influence of technical bias on selfing rates derived from inbreeding coefficients can be quantified in the case of null alleles. The occurrence of such alleles is a more serious problem in outbred populations, because they remain undetected as null homozygotes are rare. For example, if the frequency of a null allele is 0.1 in an outbred population ($S=0$), the expected frequency of null homozygotes is 0.01, and they go unnoticed given the usual sample size of about 30–40 individuals per population. From the simple model presented in Supplementary Appendix 4, it can be shown that the expected difference between the estimated \hat{F} and the actual value of F is about $(2-F)(1-F)p_n$ with p_n the frequency of the null allele. At low S , the expected bias

on \hat{F} is therefore approximately twice, and that on \hat{S} little less than four times, the frequency of the null allele (Figure 1). This bias can override the actual (null or small) F -value in outbred species. Note that similar biases are expected for other types of partial dominance problems (Supplementary Appendix 4). A general approximation for the bias on selfing rates is $\alpha(1-S)(2-S)/(1+\alpha(1-S))$, with α the fraction of heterozygotes read as homozygotes. The bias is of order 2α when S and α are low. For all sources of artifacts, including null alleles, the bias decreases when S increases.

The influence of genotyping errors on PAA estimates has not been sorted out quantitatively, although Ritland notes in MLTR documentation that family estimates might be seriously biased by misscoring. We illustrate the case of null alleles using a simple model developed in Supplementary Appendix 6. We consider a specific (perhaps worst case) situation in which both null homozygous mothers and homozygous offspring are discarded from the analysis. In such a situation, the selfing rate is overestimated, and when S is small, the bias is about $4p/3$ (where p is the frequency of the null allele). It decreases with increasing selfing rates (Figure 1 in Supplementary Appendix 6). Although this bias should not be neglected, it is much lower than that when S is estimated from the inbreeding coefficient. In addition, as mentioned above, null alleles result in inconsistencies between mother and offspring genotypes, and should therefore be more easily diagnosed in progeny arrays than in single-generation samples. The probability of detecting at least one critical offspring genotype (incompatible with mother genotype) rises fast with the number of offspring (n). For example, it can be shown from Table 1 in Appendix 6 that if the mother is heterozygous for two nonnull alleles this probability is about $2/3$ when S is low and $n=10$. On the whole, genotyping errors more severely limit the inbreeding-coefficient approach than the PAA.

Getting estimates of S free from technical bias, especially under the PSA, therefore seems a worthwhile goal. This is exactly what is achieved by the multilocus method of David *et al.* (2007) that is insensitive to the technical artifacts that generate apparent heterozygote deficiencies, including null alleles. This is because the estimation of g_2 is independent of F : technical biases create heterozygote deficiencies (bias in F), but do not create correlations in heterozygosity among loci because they occur independently at different loci. An illustration is provided in Figure 2. This method is particularly useful when S is low, that is when technical biases such as null alleles are more of a problem.

Methodological bias

These biases are essentially related to violation of assumptions regarding population structure and history or to selection, and have variable consequences on \hat{S} depending on the method (Table 1; Supplementary Appendix 6). Biases in \hat{S} due to population structure and history are generally positive (Table 1). Various processes enter this category: (1) The first is biparental inbreeding, but its magnitude is likely to be limited. Jarne and Auld (2006) estimated from the plant data set from Goodwillie *et al.* (2005) that biparental inbreeding

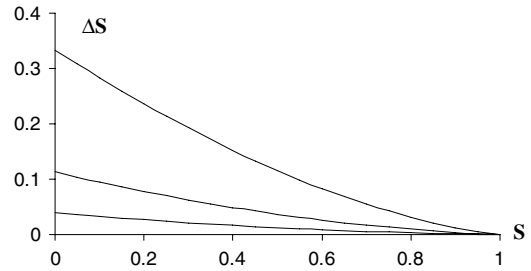


Figure 1 Difference, ΔS , between the selfing rate estimated from the inbreeding coefficient (F) and the actual selfing rate (S) as a function of S , when null alleles affect inbreeding estimates. The curves are based on equations provided in Supplementary Appendix 4. From top to bottom, the null-allele frequency is 0.1, 0.03 and 0.01, respectively.

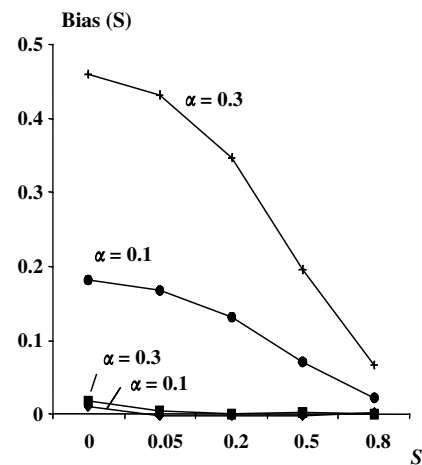


Figure 2 Bias on single-generation estimates of S using the inbreeding coefficient method averaged over loci (top two curves) and the multilocus method based on g_2 (bottom curves). Each value was obtained from 300 simulations of a sample of 100 individuals. Ten loci with gene diversity $H_e=0.8$ each were considered. Two misscoring rates were simulated, $\alpha=0.1$ and $\alpha=0.3$, with α the fraction of heterozygotes scored as homozygotes.

on average inflated apparent selfing rates by about 3%, irrespective of the true value. (2) The Wahlund effect results from mixing genetically differentiated subpopulations. Estimates of S are inflated because this process generates heterozygote deficiencies. If these populations differ in heterozygosity, the multilocus estimates obtained from the method of David *et al.* (2007) method will also be inflated. In general, the Wahlund effect is unlikely to play a significant role because sampling rarely includes strongly differentiated subpopulations. This is perhaps more questionable with the PAA, because allelic frequencies in the male gamete pool may differ from those in the female gamete pool (for example, as a consequence of pollinator behavior in plants). Such a difference can be evaluated using specific genetic analyses. (3) Deviation from inbreeding equilibrium is an issue for the PSA only, mainly when S is large and varied dramatically in the recent past. In the latter case the estimate will be between the current and past values of S . Such events can be detected using the multilocus method proposed by Enjalbert and David (2000). (4) Linkage disequilibrium is of course more of an issue

for all multilocus methods, under both the PSA and the PAA, especially at high selfing rates. Under the PAA, it leads to overestimating S , and underestimating biparental inbreeding (Hedrick and Ritland, 1990). Its influence is less one-sided under the method of David *et al.* (2007) method. Whatsoever linkage equilibrium should be checked independently (see for example, Rousset and Raymond, 1997). The influence of asexual reproduction in those species able to both self-fertilize and reproduce clonally has not been sorted out for the multilocus PSA (David *et al.*, 2007), and can formally be taken into account in the PAA. On the other hand, it might slightly lower estimates of the inbreeding coefficient (Prugnolle *et al.*, 2005). The problem can partially be alleviated by discarding copies of multilocus genotypes represented more than once.

Selection is a serious issue for all methods, as inbreeding depression is a prevalent feature of most species, including highly inbred ones (Husband and Schemske, 1996). The inbreeding coefficient at conception, corresponding to the primary selfing rate (value at fertilization before any selection), differs from that in adults as a consequence of inbreeding depression (see Supplementary Appendix 3). Whatever the method, estimates derived from adult individuals might therefore underestimate the primary selfing rate, especially in outcrossing species in which high inbreeding depression is expected early in ontogeny (Doums *et al.*, 1996; Husband and Schemske, 1996). In this respect, it is worth conducting PSAs on individuals as young as possible. The same problem occurs in principle with the PAA, but is often mitigated in practice because progenies are genotyped as either seeds, or juveniles. Estimates can be corrected using independent estimates of inbreeding depression, accounting for selection up to the stages at which S is estimated. Another way out is the methodology proposed by Ritland (1990a): sampling at two stages (say seedlings and adults) allows jointly estimating F , S and inbreeding depression expressed between the two stages (Supplementary Appendix 3).

Computational problems can be encountered when the selfing rate is numerically estimated using maximum likelihood (multilocus PSA and PAA). Convergence is indeed not granted, or might be slow. Ritland (1990b, 2002) used two methods for maximizing likelihoods partly as a solution to this issue. On the other hand, estimating likelihoods allows building tests among competing models, a desirable quality for building a hypothesis-testing framework.

Comparing estimates derived from different methods

We have described several methods for quantifying inbreeding in populations, and one might be interested in comparing parameter estimates across methods. The main reason is that estimates from the PSA and the PAA represent different views of inbreeding, the first assesses inbreeding over several generations, while the second is more a 'here and now' estimate. Another more practical reason is that the maternal F is estimated together with S among offspring in the PAA.

We have already seen that estimates of S derived from the inbreeding-coefficient and multilocus (PSA)

approaches might fruitfully be compared, for example to uncover potential technical pitfalls. This point is illustrated in David *et al.* (2007): these authors provide the striking example of an allozymic data set based on 12 populations of the freshwater snail *Physa acuta*. F -estimates suggested selfing rates equal or higher than 0.5, while the multilocus approach detected deviation from random mating in two populations only. This last result is much more in line with what is known from PAA and population structure analysis based on microsatellites (Henry *et al.*, 2005). A more systematic comparison of results from the two approaches would be worthwhile, and might even provide insights on the magnitude of technical problems in species where the inbreeding rate is known from other approaches.

F -estimates of S have also been compared to PAA estimates (Jarne and Auld, 2006). It should be realized that a one-to-one relationship is not expected because estimates are not derived from the same life-history stage. Adult mothers should have lower apparent selfing rate than their offspring because of inbreeding depression (Supplementary Appendix 3). In the quite restricted animal data set, this difference was not observed, although inbreeding depression is usually strong in outbreeding species, suggesting that maternal F -estimates were overestimated due to technical problems (Jarne and Auld, 2006). The same relationship in plants is more in line with expectations, and consistent with some inbreeding depression at low selfing rates and more limited technical bias. However, one-third of F -estimates are negative, a surprising result with no obvious interpretation.

Recommendations

We propose simple recommendations for future studies (see also Table 1). These are mostly 'rules of thumb' based on the arguments developed above and our experience, providing estimates of selfing rates with a precision of, say, 10%. For more specific goals (for example, testing for difference among families or subpopulations), appropriate designs should be built.

(1) Markers: Preferentially use microsatellites, and several loci. A good target is $LH=5$, where L is the number of loci and H their average genetic diversity. (2) Sampling: $n=30-50$ individuals is a minimum for the PSA; larger numbers are required in the PAA, because sampling is structured in families (typical numbers are 20 families of 5-10 offspring, or more if maternal genotypes have to be inferred). There certainly is a trade off between the number of families and the number of offspring per family depending on whether the focus is on within-family or among-family variation. Individuals should be sampled as young as possible if one is interested in primary selfing rates. (3) Analysis: use both inbreeding coefficients and the multilocus method (RMES) when analyzing population samples, and multilocus maximum likelihood (MLTR) for the PAA. (4) Data interpretation should carefully consider how technical biases and violation of assumptions might interfere with the estimation of S . In this respect, several complementary analyses provide invaluable help, whatever the approach. These include evaluating genotyping repeatability, comparing single-locus and multilocus estimates, comparing different age-classes or estimating inbreeding

depression, documenting the population genetic context, especially population substructure and gametic disequilibria, and jointly using the PAA and PSA whenever possible.

Conclusion and perspectives

Estimates of S can be derived from a variety of approaches, and it is certainly worth combining them because they document various aspects of the selfing rate, which is indeed a variable, dynamic parameter. No method is devoid of assumptions, and the interpretation of results from other analyses (for example, estimating inbreeding depression, analyzing population genetic substructure) has proved valuable in several species. Fortunately the scope of the PAA has been continuously enlarged over the years (Ritland, 2002), and the approach of David *et al.* (2007) renders the PSA as robust as the PAA. However, there is room for theoretical improvements, for example to account for partial dominance in the PAA (see Supplementary Appendix 6) or to extend its scope to other biological situations (see MLTR documentation). A significant effort should also be devoted to reducing technical biases as much as possible, as they arguably are the most important threat to the validity of selfing estimates.

Finally, the benefit of increasingly refining methods for estimating selfing is not to slowly converge toward the true selfing rate of a species, for there is no such thing. On the contrary, removing estimation bias and error should allow researchers to focus on what selfing really is: a variable trait and the product of a complex interaction between genotypes and environment (Brown *et al.*, 1989), which constitutes a worthy subject of study on its own. Even flowers on the same plant can display different selfing rates, and the selfing rate may also vary during the lifetime of an individual (for example, along the reproductive season in a snail). Properly quantifying fine-grained variation in S is a prerequisite to understanding how much of this variation reflects adaptation versus environmental or developmental constraints. A promising avenue for further methodological developments is to refine the decomposition of the variance in selfing rates among different levels of biological organization (populations, families, individuals or even different ramets or successive reproductive events of a single individual) and their covariance with relevant environmental variables in natural as well as in experimental settings.

Acknowledgements

We thank J Auld, JS Escobar and two anonymous referees for comments on the article, and K Ritland for discussions. This work was supported by funds from the French Centre National de la Recherche Scientifique.

References

- Avise JC (2000). *Phylogeography*. Harvard University Press: Cambridge, MA.
- Ayres KL, Balding DJ (1998). Measuring departures from Hardy-Weinberg: a Markov chain Monte-Carlo method for estimating the inbreeding coefficient. *Heredity* **80**: 769–777.

- Barrett SCH (2003). Mating strategies in flowering plants: the outcrossing-selfing paradigm and beyond. *Philos Trans R Soc Lond B* **358**: 991–1004.
- Blouin MS (2003). DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends Ecol Evol* **18**: 503–511.
- Brown AHD (1990). Genetic characterization of plant mating systems. In: Brown AHD, Clegg MT, Kahler AL, Weir BS (eds). *Plant Population Genetics, Breeding, and Genetic Resources*. Sinauer Associates: Sunderland. pp 145–162.
- Brown AHD, Allard FW (1970). Estimation of the mating system in open-pollinated maize populations using isozyme polymorphisms. *Genetics* **66**: 133–145.
- Brown AHD, Buron JJ, Jarosz AM (1989). Isozyme analysis of plant mating systems. In: Soltis D, Soltis P (eds). *Isozymes in Plant Biology*. Dioscorides Press. pp 73–86.
- Charlesworth D (2003). Effects of inbreeding on the genetic diversity of populations. *Philos Trans R Soc Lond B* **358**: 1051–1070.
- Clegg MT, Kahler AL, Allard RW (1978). Estimation of life cycle components of selection in an experimental plant population. *Genetics* **89**: 765–792.
- Criscione CD, Blouin MS (2006). Minimal selfing, few clones, and no among-host genetic structure in a hermaphroditic parasite with asexual larval propagation. *Evolution* **60**: 553–562.
- Cruzan MB (1998). Genetic markers in plant evolutionary ecology. *Ecology* **79**: 400–412.
- Curie-Cohen M (1982). Estimates of inbreeding in a natural population: a comparison of sampling properties. *Genetics* **100**: 339–358.
- Cutter AD (2006). Nucleotide polymorphism and linkage disequilibrium in wild populations of the partial selfer *Caenorhabditis elegans*. *Genetics* **172**: 171–184.
- Dakin EE, Avise JC (2004). Microsatellite null alleles in parentage analysis. *Heredity* **93**: 504–509.
- Darwin C (1876). *The Effects of Cross and Self Fertilisation in the Vegetable Kingdom*. John Murray: London.
- David P, Perdieu M-A, Pernot A-F, Jarne P (1997). Fine-grained spatial and temporal population genetic structure in the marine bivalve *Spisula ovalis* L. *Evolution* **51**: 1318–1322.
- David P, Pujol B, Viard F, Castella E, Goudet J (2007). Reliable selfing rate estimates from imperfect population genetic data. *Mol Ecol* **16**: 2474–2487.
- Doums C, Viard F, Pernot A-F, Delay B, Jarne P (1996). Inbreeding depression, neutral polymorphism, and copulatory behavior in freshwater snails: a self-fertilization syndrome. *Evolution* **50**: 1908–1918.
- Enjalbert J, David JL (2000). Inferring recent outcrossing rates using multilocus individual heterozygosity: application to evolving wheat populations. *Genetics* **156**: 1973–1982.
- Eppley SM, Taylor PT, Jesson LK (2007). Self-fertilization in mosses: a comparison of heterozygote deficiency between species with combined versus separate sexes. *Heredity* **98**: 38–44.
- Estoup A, Angers B (1998). Microsatellites and minisatellites for molecular ecology: theoretical and empirical considerations. In: Carvalho G (eds). *Advances in Molecular Ecology*. NATO press: Amsterdam. pp 55–86.
- Excoffier L (2001). Analysis of population subdivision. In: Balding DJ, Bishop M, Cannings C (eds). *Handbook of Statistical Genetics*. John Wiley & Sons: Chichester. pp 271–307.
- Fryxell PA (1957). Mode of reproduction in higher plants. *Bot Rev* **23**: 135–233.
- Fyfe JL, Bailey NTJ (1951). Plant breeding studies in leguminous forage crops. I. Natural cross-breeding in winter beans. *J Agric Sci* **41**: 371–378.
- Goodwillie C, Kalisz S, Eckert CG (2005). The evolutionary enigma of mixed mating in plants: occurrence, theoretical explanations, and empirical evidence. *Annu Rev Ecol Evol Syst* **36**: 47–79.

- Goudet J (1995). FSTAT (Version 1.2): a computer program to calculate F-statistics. *J Hered* **86**: 485–486.
- Hedrick PW, Ritland K (1990). Gametic disequilibrium and multilocus estimation of selfing rates. *Heredity* **65**: 343–347.
- Henry P-Y, Bousset L, Sourrouille P, Jarne P (2005). Partial selfing, ecological disturbance and reproductive assurance in an invasive freshwater snail. *Heredity* **95**: 428–436.
- Hoffman JI, Amos W (2005). Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. *Mol Ecol* **14**: 599–612.
- Husband BC, Schemske DW (1996). Evolution of the magnitude and timing of inbreeding depression. *Evolution* **50**: 54–70.
- Irwin RE, Strauss SY (2005). Flower color microevolution in wild radish: Evolutionary response to pollinator-mediated selection. *Am Nat* **165**: 225–237.
- Jarne P (1995). Mating system, bottlenecks and genetic polymorphism in hermaphroditic animals. *Genet Res* **65**: 193–207.
- Jarne P, Auld JR (2006). Animals mix it up too: the distribution of self-fertilization among hermaphroditic animals. *Evolution* **60**: 1816–1824.
- Jarne P, Charlesworth D (1993). The evolution of selfing rate in functionally hermaphrodite plants and animals. *Annu Rev Ecol Syst* **24**: 441–466.
- Jarne P, Lagoda PJL (1996). Microsatellites, from molecules to populations and back. *Trends Ecol Evol* **11**: 424–429.
- Jones DF (1916). Natural cross-pollination in the tomato. *Science* **43**: 509–510.
- Li CC (1955). *Population Genetics*. University of Chicago Press: Chicago.
- Murawski DA, Fleming TH, Ritland K, Hamrick JL (1994). Mating system of *Pachycereus pringlei*—an autotetraploid cactus. *Heredity* **72**: 86–94.
- Pompanon F, Bonin A, Bellemain E, Taberlet P (2005). Genotyping errors: causes, consequences and solutions. *Nat Rev Genet* **6**: 847–859.
- Prugnolle F, Roze D, Théron A, De Meeus T (2005). F-statistics under alternation of sexual and asexual reproduction: a model and data from schistosomes (platyhelminth parasites). *Mol Ecol* **14**: 1355–1365.
- Renner SS, Ricklefs RE (1995). Dioecy and its correlates in flowering plants. *Am J Bot* **82**: 596–606.
- Ritland K (1984). The effective proportion of self-fertilization with consanguineous matings in inbred populations. *Genetics* **106**: 139–152.
- Ritland K (1986). Joint maximum-likelihood-estimation of genetic and mating structure using open-pollinated progenies. *Biometrics* **42**: 25–43.
- Ritland K (1989). Correlated matings in the partial selfer *Mimulus guttatus*. *Evolution* **43**: 848–859.
- Ritland K (1990a). Inferences about inbreeding depression based on changes of the inbreeding coefficient. *Evolution* **44**: 1230–1241.
- Ritland K (1990b). A series of FORTRAN computer programs for estimating plant mating systems. *J Hered* **81**: 235–237.
- Ritland K (2002). Extensions of models for the estimation of mating systems using n independent loci. *Heredity* **88**: 221–228.
- Ritland K, El Kassaby YA (1985). The nature of inbreeding in a seed orchard of douglas-fir as shown by an efficient multilocus model. *Theor Appl Genet* **71**: 375–384.
- Ritland K, Jain SK (1981). A model for the estimation of outcrossing rate and gene frequencies using n independent loci. *Heredity* **47**: 35–52.
- Ronfort JL, Jenczewski E, Bataillon T, Rousset F (1998). Analysis of population structure in autotetraploid species. *Genetics* **150**: 921–930.
- Rousset F (2001). Inferences from spatial population genetics. In: Balding DJ, Bishop M, Cannings C (eds). *Handbook of Statistical Genetics*. John Wiley & Sons: Chichester. pp 239–269.
- Rousset F, Raymond M (1997). Statistical analyses of population genetic data: new tools, old concepts. *Trends Ecol Evol* **12**: 313–317.
- Schemske DW, Lande R (1985). The evolution of self-fertilization and inbreeding depression in plants. II. Empirical observations. *Evolution* **39**: 41–52.
- Schoen DJ, Clegg MT (1986). Monte-Carlo studies of plant mating system estimation models—the one-pollen parent and mixed mating models. *Genetics* **112**: 927–945.
- Selkoe KA, Toonen RJ (2006). Microsatellite for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecol Lett* **9**: 615–629.
- Shaw DV, Kahler AL, Allard RW (1981). A multilocus estimator of mating system parameters in plant populations. *Proc Natl Acad Sci USA* **78**: 1298–1302.
- Thompson SL, Ritland K (2006). A novel mating system analysis for modes of self-oriented mating applied to diploid and polyploid arctic Easter daisies (*Townsendia hookeri*). *Heredity* **97**: 119–126.
- Uyenoyama MK, Holsinger KE, Waller DM (1993). Ecological and genetic factors directing the evolution of self-fertilization. *Oxf Surv Evol Biol* **9**: 327–381.
- Vianey-Liaud M (1997). La reproduction chez un mollusque hermaphrodite simultané, la planorbe *Biomphalaria glabrata* (Say, 1818) (Gastéropode, Pulmoné). *Halictis* **27**: 67–114.
- Viard F, Brémond P, Labbo R, Justy F, Delay B, Jarne P (1996). Microsatellites and the genetics of highly selfing populations in the freshwater snail *Bulinus truncatus*. *Genetics* **142**: 1237–1347.
- Vos P, Hogers R, Bleeker M, Reijans M, Vandelee T, Hornes M *et al.* (1995). AFLP—a new technique for DNA-fingerprinting. *Nucleic Acids Res* **23**: 4407–4414.
- Weir BS, Cockerham CC (1973). Mixed self and random mating at two loci. *Genet Res* **21**: 247–262.
- Weir BS, Cockerham CC (1984). Estimating F-statistics for analysis of population structure. *Evolution* **38**: 1358–1370.

Supplementary Information accompanies the paper on Heredity website (<http://www.nature.com/hdy>)