Molecular Ecology (2007) **16**, 2474–2487 doi: 10.1111/j.1365-294X.2007.03330.x

Reliable selfing rate estimates from imperfect population genetic data

PATRICE DAVID,* BENOÎT PUJOL,*† FRÉDÉRIQUE VIARD,‡ VINCENT CASTELLA§¶ and JÉRÔME GOUDET¶

*CEFE-CNRS, UMR 5175, Montpellier & France 1919 Route de Mende, 34293 Montpellier Cedex 05, France, †Department of Plant Sciences, University of Oxford, South Parks Road, Oxford, OX1 3RB, UK, ‡«Evolution et Génétique des Populations Marines», AD2M UMR CNRS-UPMC 7144, Station Biologique de Roscoff, Place Georges-Teissier, BP74, 29682 Roscoff Cedex, France, §IUML, Rue du Bugnon 21, CH-1005 Lausanne, Switzerland, ¶DEE, Biophore, UNIL-Sorge, CH-1015 Lausanne, Switzerland

Abstract

Genotypic frequencies at codominant marker loci in population samples convey information on mating systems. A classical way to extract this information is to measure heterozygote deficiencies (F_{IS}) and obtain the selfing rate s from $F_{IS} = s/(2 - s)$, assuming inbreeding equilibrium. A major drawback is that heterozygote deficiencies are often present without selfing, owing largely to technical artefacts such as null alleles or partial dominance. We show here that, in the absence of gametic disequilibrium, the multilocus structure can be used to derive estimates of s independent of F_{IS} and free of technical biases. Their statistical power and precision are comparable to those of F_{IS}, although they are sensitive to certain types of gametic disequilibria, a bias shared with progeny-array methods but not F_{IS}. We analyse four real data sets spanning a range of mating systems. In two examples, we obtain $s = 0$ despite positive F_{IS}, strongly suggesting that the latter are artefactual. In the remaining examples, all estimates are consistent. All the computations have been implemented in a open-access and user-friendly software called RMES (robust multilocus estimate of selfing) available at http://ftp.cefe.cnrs.fr, and can be used on any multilocus data. Being able to extract the reliable information from imperfect data, our method opens the way to make use of the ever-growing number of published population genetic studies, in addition to the more demanding progeny-array approaches, to investigate selfing rates.

Keywords: heterozygote deficiency, identity disequilibrium, inbreeding, mating system, microsatellite, null alleles

Received 31 October 2006; revision accepted 19 February 2007

Introduction

Estimating selfing rates is a basic step in many population studies, especially in plants (Schemske & Lande 1985; Jarne & Charlesworth 1993). Since the advent of molecular markers, increasingly sophisticated methods to infer selfing rates from molecular data have been developed (Fyfe & Bailey 1951; Brown & Allard 1970; Ritland & Jain 1981; Shaw *et al*. 1981; Ritland 2002). The heterozygosity of the typed individuals (with or without reference to their parent's genotype) is the main source of information on their inbreeding status in all these methods. Two principal ways to infer selfing rates from molecular data, are based on, respectively, the estimation of fixation indexes (F_{IS}) in population samples and the use of progeny arrays (Fyfe & Bailey 1951; Ritland & Jain 1981). These two methods are complementary: progeny arrays convey detailed information on the mating system of the last generation, including among-family variance in selfing and paternity correlations. F_{IS} is a more integrative measure because it integrates the effect of selfing together with any other form of biparental inbreeding over several generations. In plants, a vast majority of estimates of selfing rates are now drawn from progeny arrays (e.g. in the large data set compiled by Schemske & Lande 1985; updated in Goodwillie *et al*. 2005), suggesting that geneticists are not as confident in indirect, F_{IS}-based estimates as they are in the direct method of

Correspondence: P. David, Fax: (33) 67 41 21 38; E-mail: patrice.david@cefe.cnrs.fr

© 2007 Centre National de la Recherche Scientifique
Journal compilation © 2007 Blackwell Publishing Ltd

progeny arrays. Yet, population data sets are commonly found in the literature especially for animals in which progenies are not always easy to get: as recently reviewed by Jarne & Auld (2006) nearly all animal mating-system studies are based on F_{IS}. Moreover, the ease with which multilocus F_{IS} data can be obtained makes them more appropriate for large-scale comparisons. It thus seems a pity to discard such potential amounts of information.

One of the main reasons for being cautious with F_{IS}-based estimates of selfing certainly lies with typing artefacts (Hoffman & Amos 2005; Pompanon *et al*. 2005). Even dioecious species often show large heterozygote deficiencies, not nearly compatible with what would look like reasonable inbreeding levels for the species (see, e.g. Zouros & Foltz 1984). Often, F_{IS} values are quite different among loci, which no form of systematic inbreeding can account for. For example, F_{IS} varies from –0.01 to 0.40 in Gaffney *et al*. (1990), in which sampling variance is small owing to the large sample size $N = 1906$ and large observed heterozygosities. Of course, population substructure or homogamy could also produce positive F_{IS}. However, after critical examination of all hypotheses, scoring artefacts often remain the only realistic explanation (David *et al*. 1997b; Bonin *et al*. 2004).

Crucially, for the estimation of F_{IS}, scoring should not bias the frequency of heterozygous genotypes. Unfortunately, one does expect a negative bias on heterozygosity, given that null alleles and partial dominance, two of the main sources of errors, both result in misscoring heterozygotes as homozygotes. This problem concerns both allozymes and (polymerase chain reaction) PCR-based markers, in which nonamplified alleles (null alleles) and asymmetric competition between sequences during amplification (short-allele dominance or large allele drop-out) are a recurrent problem (Wattier *et al*. 1998; Bonin *et al*. 2004; Hoffman & Amos 2005; Pompanon *et al*. 2005). Of course, a number of methods exist to detect such artefacts (Brookfield 1996; Van Oosterhout *et al*. 2004; Björklund 2005), and to a certain extent to correct them (starting with simple common sense: a conservative elimination of loci with extreme heterozygote deficiency). However, most of these methods address only one specific type of misscoring, and were designed to correct allele frequencies rather than heterozygosity, assuming either Hardy–Weinberg equilibrium (Brookfield 1996), or a known inbreeding coefficient (Van Oosterhout *et al*. 2006). Thus, it is extremely difficult to extract reliable F_{IS} estimates (or, in fact, any other statistic dependent on observed heterozygosity) corrected for typing errors, simply because it is not possible to estimate with reasonable precision how many of the missing heterozygotes are due to inbreeding and to artefacts, respectively. Averaging F_{IS} over several loci (as usually done) obviously averages the bias rather than suppresses it.

Is there a way to estimate selfing rates in population samples at equilibrium without using F_{IS}? The multilocus structure of the sample in principle offers an independent source of information. Indeed, partial selfing not only creates heterozygote deficiencies, it also generates identity disequilibria, i.e. correlations in heterozygosity among different loci (Bennett & Binet 1956; Weir & Cockerham 1973). Identity disequilibria, defined as relative excess in doubly heterozygous genotypes (Weir & Cockerham 1973), can be measured for all pairs of loci (or, even more, pairs of alleles at different loci), which sums up to a large number of parameters. However, partial selfing generates identical disequilibria at all pairs of loci and alleles (Weir & Cockerham 1973), allowing to construct multilocus statistics directly linked to the selfing rate, independent of F_{IS} and jointly estimable on an arbitrary number of loci. Unless scoring artefacts are correlated across loci within an individual, such statistics will remain unaffected whatever the bias on F_{IS}. Another way to put it is that partial selfing does not only modify the average inbreeding level of individuals (and thus, the average F_{IS} per locus) but also the variance in inbreeding among individuals (Bierne *et al*. 2000): some are produced by outcrossing, others by one or more generations of selfing. This variance in turn produces an excess of multihomozygous or multiheterozygous genotypes compared to random assortment of single-locus genotypes, i.e. a disequilibrium in the distribution of multiple-locus heterozygosity (MLH) that can be used to infer the selfing rate, whatever the average observed heterozygosity. A similar idea was already put forward by Enjalbert & David (2000), although their method did make use of the average heterozygosity (and therefore was not free of scoring artefacts), and they focused on recent temporal changes in selfing rates in domesticated populations rather than equilibrium populations.

We here describe a general model providing a method to extract selfing rates from the multilocus correlation structure of a population sample. Two related methods are proposed, one based on a point estimate of the second-order heterozygosity disequilibrium (see below), and the other based on maximum-likelihood of the whole distribution. The statistical behaviour of both methods is evaluated in terms of bias, variance, and sensitivity to misscoring. Their use is illustrated on four plant and animal data sets spanning a range of selfing rates.

The model

The effects of the selfing rate and scoring biases on the distribution of multilocus apparent heterozygosity

We consider a sample of individuals from a single population at inbreeding equilibrium, in which heterozygosity is recorded at a set of L loci. The apparent heterozygosity of an individual at locus i ($H_i = 0$ if homozygous, 1 if heterozygous) may differ from its true heterozygosity (h_i)

© 2007 Centre National de la Recherche Scientifique
Journal compilation © 2007 Blackwell Publishing Ltd

because of scoring artefacts. We assume that artefacts lead to scoring true heterozygotes as homozygotes with a probability x_i that varies among loci. x_i may also take negative values (minimum: $1 - 1/E(h_i)$) if artefacts have the opposite effects. We also assume that misscorings occur independently at different loci of the same individual. This assumption may be violated if, for example, misscoring is due to the poor quality of a single DNA extract used to type several loci; however, in practice, such samples are likely to be detected (some loci may not amplify at all) and discarded from the analysis.

The effect of selfing and other forms of systematic inbreeding can be captured by the distribution of f, a variable that represents the probability that two genes at a locus in an individual will be identical by descent (Weir *et al.* 1980). The mean of this distribution, $E(f)$, usually noted F_{IS}, is well-known as $E(f) = s/(2 - s)$ in an infinite population at inbreeding equilibrium with selfing rate s. This mean f (and hence the selfing rate) can be directly estimated by heterozygote deficiencies at a neutral marker (Robertson & Hill 1984; Weir & Cockerham 1984). However, it will necessarily be biased (usually upwards) in the presence of misscorings. On the other hand, partial inbreeding also creates heterogeneity in f within the population, which can be captured by the variance and higher moments of the distribution of f. Because individuals with high f tend to be homozygous at all loci at the same time (Bennett & Binet 1956,Weir & Cockerham 1973), the distribution of multiple-locus heterozygosity (MLH) will deviate from the expectation under random assortment of single-locus heterozygosities. With two loci, this covariance in heterozygosity is usually quantified through the identity disequilibrium (see Weir & Cockerham 1973). We will here generalize this approach to k loci ($k > 0$), defining the kth-order heterozygosity disequilibrium g_k as the relative excess of genotypes heterozygous at k loci, compared to independent assortment:

$$E(h_1 h_2 \ldots h_k) = E(h_1)E(h_2) \ldots E(h_k)\,(1 + g_k). \qquad \text{(eqn 1)}$$

Below, we will show that g_k is independent of the choice of the k loci when within-population inbreeding is the sole source of deviation from Hardy–Weinberg equilibrium. This can be done by conditioning on F values. Noting that the probability of being heterozygous is reduced by a proportion f at each locus, and that within a class of individuals that have the same value of f (hereafter, an inbreeding level), there is no correlation in heterozygosity among loci:

$$\begin{aligned} E(h_1 h_2 \ldots h_k) &= \sum_{0 \le f \le 1} p(f)E(h_1 \mid f)E(h_2 \mid f) \ldots E(h_k \mid f) \\ &= D_1 D_2 \ldots D_k E((1 - f)^k) \end{aligned} \qquad \text{(eqn 2)}$$

where D_i is gene diversity at locus i. Given that $E(h_i) = (1 - E(f))D_i$, the expression of g_k only depends on the characteristics of the distribution of f:

$$g_k = \frac{E((1 - f)^k)}{[E(1 - f)]^k} - 1 \qquad \text{(eqn 3)}$$

The values of $g_1 \ldots g_k$ can be easily computed using classical methods under various schemes of partial inbreeding (see Weir *et al.* 1980). In the most classical scheme, i.e. partial selfing at rate s in an infinite population, we obtain

$$g_k = \frac{1 - s}{(1 - s/2^k)\left[1 - \left(\dfrac{s}{2 - s}\right)\right]^k} - 1 \qquad \text{(eqn 4)}$$

Under partial selfing, the g_k values vary between 0 and $+\infty$, and increase with the number of loci k, meaning that the relative excess in multiheterozygotes (compared to random assortment across loci) is enhanced.

Let us now evaluate the effect of misscoring. At each locus i, misscoring will decrease apparent heterozygosity by a factor $(1 - x_i)$ compared to true heterozygosity. Note that our goal here is not to evaluate the extent and kind of technical errors, but rather to derive a method insensitive to them in order to estimate s. Nonetheless, x values can be obtained as a by-product of our analyses (see below). The estimate of $E(f)$ (or F_{IS}) based on heterozygote deficiency, will be inflated by an amount proportional to $x_i\,(1 - f)$ (see Appendix I). Unlike sampling error, this bias is independent of sample size. Interestingly, provided scoring artefacts do not occur in a correlated fashion across loci, the kth-order heterozygosity disequilibrium is not affected, for $k > 1$:

$$\begin{aligned} E(H_1 H_2 \ldots H_k) &= (1 - x_1) \ldots (1 - x_k)E(h_1 h_2 \ldots h_k) \\ &= E(H_1)E(H_2) \ldots E(H_k)(1 + g_k) \end{aligned} \qquad \text{(eqn 5)}$$

In other terms, heterozygosity disequilibria can be equivalently computed on apparent rather than on true heterozygosities. A simple way to estimate selfing rates without technical bias is therefore to estimate one of these coefficients (the simplest being g_2) and invert the formula given in equation 4. Technical artefacts tend to decrease the mean apparent heterozygosity at each locus but leave the covariance unaffected because they occur independently at each locus. This is why our estimate, based only on covariances, is robust to artefacts. In the same way, it is also possible to derive a maximum-likelihood estimate of s based on the complete distribution of apparent heterozygosities, without reference to true heterozygosities or true heterozygote deficiencies. Below, we first derive a minimum-bias estimator of g_2 and its statistical properties then detail the maximum-likelihood method.

Estimation of second-order heterozygosity disequilibrium and selfing rate using multilocus data

As shown above, g_2 is not dependent on the pair of loci chosen in equation 1. In order to extract a maximum amount

© 2007 Centre National de la Recherche Scientifique
Journal compilation © 2007 Blackwell Publishing Ltd

of information from a multilocus data set (L loci), the estimator must combine information from all pairs of loci. A simple way is to estimate the increase in variance in the number of heterozygous loci per individual relative to random-assortment of single-locus heterozygosities:

$$\begin{aligned}\mathrm{Var}\left(\sum_{i=1}^{L} H_i\right) &= \sum_{i=1}^{L}\mathrm{Var}(H_i) + 2\sum_{i=1}^{L}\sum_{j=i+1}^{L}\mathrm{Cov}(H_iH_j) \\ &= \sum_{i=1}^{L}\mathrm{Var}(H_i) + 2g_2\sum_{i=1}^{L}\sum_{j=i+1}^{L}E(H_i)E(H_j)\end{aligned} \quad \text{(eqn 6)}$$

From which we obtain immediately

$$g_2 = \frac{\mathrm{Var}\left(\sum_{i=1}^{L} H_i\right) - \sum_{i=1}^{L}\mathrm{Var}(H_i)}{2\sum_{i=1}^{L}\sum_{j=i+1}^{L}E(H_i)E(H_j)} \quad \text{(eqn 7)}$$

Taking the ratio between unbiased estimates of the numerator and denominator in equation 7 yields a low-bias estimate of g_2

$$\hat{g}_2 = \frac{\sum_{i=1}^{L}\sum_{j\neq i}\sum_{k=1}^{N} H_{ik}H_{jk}}{(N-1)^{-1}\sum_{i=1}^{L}\sum_{j\neq i}\sum_{k_1=1}^{N}\sum_{k_2\neq k_1} H_{ik_1}H_{jk_2}} \quad \text{(eqn 8)}$$

where H_{ik} is the apparent heterozygosity of individual k at locus i, L is the number of loci and N is sample size. The derivation of this estimate, and approximations for its bias and sampling variance are given in the Supplementary material, Appendix S1. A modified formula to account for missing data (i.e. unknown values for H at some loci for some individuals) is also given in this appendix.

From this, we can obtain an estimate of the selfing rate s by inverting equation 4 and taking the only root < 1:

$$\hat{s}_{g_2} = \frac{1 + 5\hat{g}_2 - \sqrt{1 + 10\hat{g}_2 + 9\hat{g}_2^2}}{2\hat{g}_2} \quad \text{(eqn 9)}$$

For small values of $\hat{g}_2$ (and s), this yields approximately $4\hat{g}_2$ (note that the selfing rate has to be set to zero if $\hat{g}_2$ turns out to be negative). The bias and variance of this estimator of s can be deduced from those of $\hat{g}_2$ (see Supplementary material, Appendix S1). Testing whether the two estimates significantly exceed zero can be done by chi-square tests (Yang 2000) which are asymptotically accurate for large samples, or simply (requiring no assumption) by generating the null distribution of each estimate under the hypothesis of no heterozygosity disequilibrium. To this end, single-locus genotypes are randomly redistributed among individuals to create a large number (say 1000) of pseudosamples from which $\hat{g}_2$ and $\hat{s}$ are estimated. The observed values are compared to null distributions to compute the P value for the hypothesis $g_2 = s = 0$. Moreover, as a by-product of our method, the rates of misscoring (x_i) can be inferred as the relative difference between the predicted and observed numbers of heterozygous genotypes, i.e. $x_i = (F_{is,\,locus\,i} - f)/(1-f)$, where $f = \hat{s}/(2-\hat{s})$.

Joint estimation of s *and apparent heterozygosity by maximum likelihood*

Several methods have already been designed to estimate s by directly maximizing the likelihood of observed single-locus (Hill *et al*. 1995; Ayres & Balding 1997) or multiple-locus (Enjalbert & David 2000) genotypes. All incorporate the deviations from Hardy–Weinberg proportions, and will therefore be affected by scoring bias. A more robust estimate will be obtained by maximizing the likelihood of the observed distribution of multilocus apparent heterozygosity, conditional on the observed mean apparent heterozygosity at each locus, without having to take into account the comparison between apparent and expected heterozygosities. This also has the advantage that the number of parameters to be estimated (in addition to the selfing rate) is considerably reduced, giving more stable estimates. Instead of estimating the frequencies of all alleles at all loci, we need to estimate only one parameter per locus: the mean apparent heterozygosity μ_i, or, equivalently the apparent heterozygosity expected under Hardy–Weinberg $d_i = (1 - x_i)D_i$, which is related to μ_i and s by $\mu_i = d_i(1 - s/(2-s))$. Our likelihood formula is therefore a function of d_i, s and multilocus genotypes, coded for each individual j as a vector of apparent heterozygosities (H_{1j}, H_{2j} … H_{Lj}) without reference to allelic states. Our method differs from other methods in that neither heterozygosities, nor gene diversities, nor allele frequencies are estimated or assumed to be known without error, as we rely only on apparent heterozygosities. Apart from s, we estimate only L 'nuisance' parameters, i.e. the μ_i (or d_i), all of which have simple binomial distributions in the sample.

To compute the likelihood of a genotype j given a set of parameters (s, $\boldsymbol{d}$), with $\boldsymbol{d} = (d_1, d_2 \ldots d_L)$, we first compute its conditional likelihood given a fixed inbreeding level f:

$$L_j(s, \boldsymbol{d}\,|\,f) = \prod_{i=1}^{L}(d_i(1-f))^{H_{ij}}(1 - d_i(1-f))^{1-H_{ij}} \quad \text{(eqn 10)}$$

We then integrate over all possible inbreeding levels, which in this case represent different numbers of generations of selfing in the pedigree since the last outcrossing event. This yields:

$$L_j(s, \boldsymbol{d}) = \sum_{t=0}^{\infty} s^t(1-s)\prod_{i=1}^{L}\left(\frac{d_i}{2^t}\right)^{H_{ij}}\left(1 - \frac{d_i}{2^t}\right)^{1-H_{ij}} \quad \text{(eqn 11)}$$

© 2007 Centre National de la Recherche Scientifique
Journal compilation © 2007 Blackwell Publishing Ltd

which can be multiplied over individuals ($j = 1$ to N) to get the likelihood of the sample. In practice, it is impossible to sum over an infinity of terms ($t = 1$ to ∞ in equation 11). A good approximation is obtained by closing the sum at some level $t_{\max}$ and equating to zero the probability of being heterozygous after $t_{\max}$ or more generations of selfing (Enjalbert & David 2000). For reasonable values of $t_{\max}$ (say 20 or more), this is a good approximation which turns out not to be too costly in computing time. The approximate likelihood is then:

$$L_j(s, \boldsymbol{d}) \approx \sum_{t=0}^{t_{\max}} s^t (1-s) \prod_{i=1}^{L} \left(\frac{d_i}{2^t}\right)^{H_{ij}} \left(1-\frac{d_i}{2^t}\right)^{1-H_{ij}} + s^{1+t_{\max}} \prod_{i=1}^{L} (1-H_{ij}) \quad \text{(eqn 12)}$$

The overall log-likelihood can be maximized numerically using a steepest-ascent algorithm and starting values given by equation 9 for s and by $(\sum_{j=1}^{N} H_{ij}/N)$ for μ_i. A good practice is to include some random noise around the starting values and run the algorithms several times to check that it has not been trapped in local optima. Confidence intervals can be determined by the profile-likelihood method. The CI95 bounds are obtained by varying s, keeping the μ_i constant, until the likelihood is decreased by the threshold value 1.95 (McCullagh & Nelder 1989). The μ_i rather than d_i are kept constant because the estimates of d_i are expected to be positively correlated with those of s (likelihood maximization tends to automatically compensate an increase in s by an increase in d so as to preserve μ_i close to the sample mean heterozygosity). Hypothesis tests are performed by likelihood-ratio (McCullagh & Nelder 1989). Various hypotheses can be tested. $s = 0$ within one population is tested by comparing the deviance [= –2 Log(Likelihood)] of the no-inbreeding model with the model with unconstrained s. The change in deviance is compared to a χ^2 with one d.f. The equality of s among different populations can be tested by comparing a model with a single value of s for all populations (allowing a different value of $\boldsymbol{d}$ for each) to a model where s is estimated independently in all populations. The change in deviance is then compared to a χ^2 with $N_{pop} - 1$ d.f., where N_{pop} is the number of populations. A Windows-compatible program called RMES (robust multilocus estimation of selfing rates) performing these operations is available in http://ftp.cefe.cnrs.fr/.

Simulations and data analyses

The properties of the above estimates were tested both by simulation and on real data. Simulations were performed by generating random arrays of genotypes coded in 0 (homozygotes) or 1 (heterozygotes) rather than by explicitly drawing allele identities, as the information of allele identities was never used. We simulated various parameter combinations with respect to (i) design parameters (sample size, number of loci), (ii) locus-specific parameters (genetic diversity D_i, misscoring rates x_i), and (iii) selfing rate s. Each multilocus genotype was generated as follows. First, the number of selfing generations t since the last outcrossing event was drawn in a geometric distribution, with probability $s^t(1 - s)$. Then, the apparent heterozygosity at each locus was drawn independently in a binary distribution with probability $(1 - x_i)D_i/2^t$. All individuals in the sample were generated independently, and the resulting matrix used to compute $\hat{g}_2$ and $\hat{s}$.

Several published data sets on different species and different marker types (allozymes and microsatellites) spanning a presumably wide range of values of selfing rates were used to illustrate our method. This comprises (i) several African populations of *Bulinus truncatus*, a freshwater snail with high selfing rate (data from Viard *et al.* 1997b); (ii) one population of cassava *Manihot esculenta* (Pujol *et al.* 2005), a monoecious plant with presumably moderate selfing rate; (iii) one population of marine bivalves *Spisula ovalis* (data from David *et al.* 1997b), a dioecious species that cannot self-fertilize; and (iv) several populations of the hermaphroditic freshwater snail *Physa acuta*, from Switzerland. This last data set has not been previously published; its interest, with respect to our method, lies in the fact that F_{IS} values suggest high selfing rates (0.5 or higher) while *P. acuta* is traditionally known as a preferential outcrosser, in other European populations (Jarne *et al.* 2000; Henry *et al.* 2005). These four studies include various number of loci and numbers of populations, typed for allozymes (for *S. ovalis* and *P. acuta*) or microsatellites (*B. truncatus* and *M. esculenta*). The details are given in corresponding references, except for *P. acuta*. For this species, six populations were sampled from lake Geneva in Switzerland (Versoix, Rolle, Vidy, Vevey and Thonon) and six others from ponds nearby (Marion, Eysins, Longirod, St Livres, Vufflens, Corcelles and Yverdon), and loci are Esterase, Nucleoside-phosphorylase, Peptidase A, B, C and D typed using standard methods from Pasteur *et al.* (1987). For all species, Weir & Cockerham's (1984) estimates of F_{IS} were calculated for all populations using GENETIX 4.05 (Belkhir *et al.* 2004), and tested by permutation tests. For comparison, estimates of selfing rates (s) were obtained using three methods: (i) based on F_{IS} [$s = 2F_{IS}/(1 + F_{IS})$], (ii) based on $\hat{g}_2$, (iii) by maximum likelihood. Using scoring-error detection software such as MICRO-CHECKER (Van Oosterhout *et al.* 2004) was unfortunately not possible in the microsatellite data, because we know that the two species concerned (cassava and *B. truncatus*) are partial selfers (or even, predominant selfers in the case of *B. truncatus*), while the software requires Hardy–Weinberg equilibrium.

In order to visualize heterozygosity disequilibria, it is also useful to plot the relative difference between observed multilocus heterozygosity (MLH) and expected MLH under random assortment of single-locus genotypes. The

© 2007 Centre National de la Recherche Scientifique
Journal compilation © 2007 Blackwell Publishing Ltd

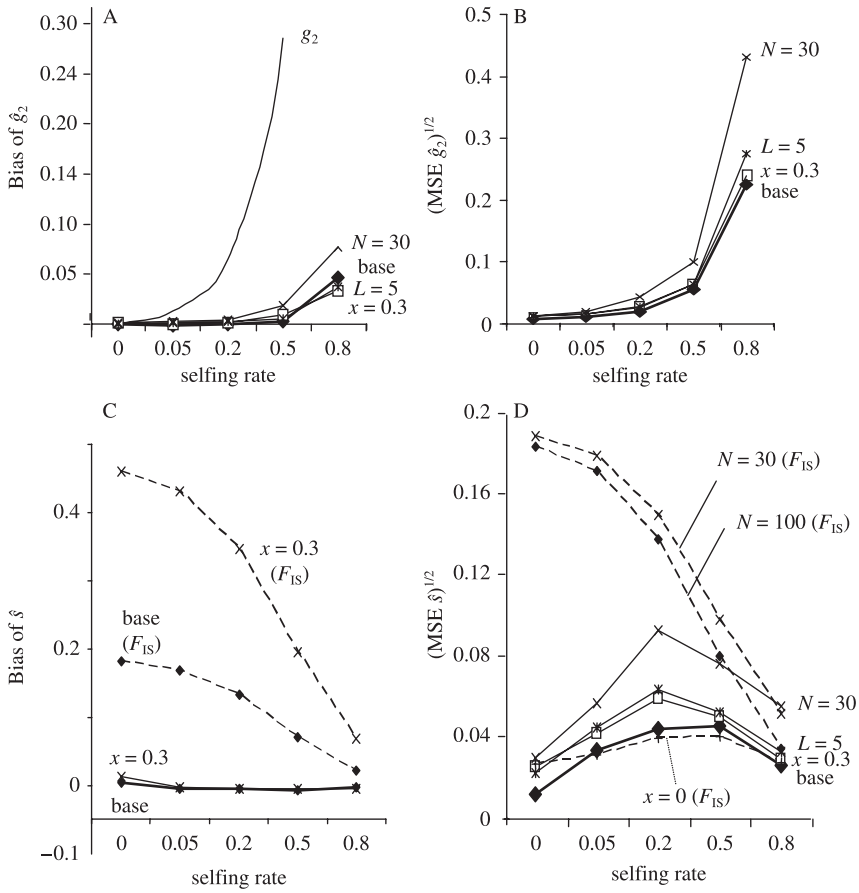

Fig. 1 Biases and mean squared error (MSE = bias^2 + sampling variance) of estimates of g_2 and s, averaged over 300 simulations. Five different selfing rates were used. The basic set of parameters ('base', filled diamonds and thick lines) is as follows: sample size $N = 100$, number of loci $L = 10$, misscoring rate $x = 0.1$, gene diversity $D = 0.8$. Other sets differ from the base by only one parameter, as labelled. (A) Bias in $\hat{g}_2$. For comparison, the values of g_2 are given for each selfing rate, except for $s = 0.8$ ($g_2 = 1.286$, out of scale). (B) MSE of $\hat{g}_2$. (C) Bottom graphs: Bias of $\hat{s}$ estimated by maximum likelihood. All graphs are visually indistinguishable, so only two were represented (base and $x = 0.3$). Estimates based on $\hat{g}_2$ give very similar results and have also been omitted. The broken lines represent the biases on selfing rates estimated using F_{IS}. (D) MSE of $\hat{s}$ (maximum likelihood method). Again, estimates based on $\hat{g}_2$ yield very similar graphs and have been omitted. For comparison, MSE of F_{IS}-based estimates are given (broken lines) for $N = 100$ (base) and $N = 30$, as well as $x = 0$ (without misscoring).

expected effect of selfing is to give this graph a U shape because it generates an excess of multihomozygous and multiheterozygous combinations, and a deficiency in intermediates. The expected distribution can be constructed very simply by recursion on the number of loci. Say $\hat{\mu}_i$ is the average apparent heterozygosity at locus i, and $p_{k,l}$ is the expected frequency of individuals with k loci heterozygous among loci 1, 2 … l. Then, for $0 < k \leq l + 1$

$$p_{k,l+1} = p_{k-1,l}\hat{\mu}_{l+1} + p_{k,l}(1 - \hat{\mu}_{l+1}) \qquad \text{(eqn 13)}$$

Starting from $p_{0,l} = \prod_{k=1}^{l}(1 - \hat{\mu}_k)$, one can generate all frequencies $p_{k,L}$ by sequentially adding all loci until $l = L$. Similar expected distributions can be obtained assuming heterozygosity disequilibria created by any hypothetical selfing rate s, instead of independent assortment of single-locus genotypes. We first compute frequencies $p_{k,L}|g$ conditional on a given number g of successive generations of selfing in the pedigree, then average over all g, weighted by their frequency $s^g(1 - s)$, using the same method as in equation 12 to approximate the infinite sum by a finite one. For any g, $p_{k,L}|g$ can be found using the same recursion system as in equation 13, replacing $\hat{\mu}_k$ by $\hat{\mu}_k/[2^g(1 - s/(2 - s))]$.

Results

The bias and mean squared error of $\hat{g}_2$ and $\hat{s}$ are given in Fig. 1. We chose basic parameters sets to represent a typical microsatellite study (100 individuals, 10 loci, $D = 0.8$) and a moderate misscoring rate ($x = 0.1$). Alterations of this set included less loci (5) or individuals (30); less polymorphic markers ($D = 0.4$, more typical of allozyme data or highly selfing species), and higher misscoring rates ($x = 0.3$). For both g_2 and ML methods, decreasing D has roughly the same effect as increasing x (keeping $(1 - x)D$ constant); so D was not varied in Fig. 1. In all cases, the analytic formulae given in Appendix S1, Supplementary material, turned out to be accurate (data not shown) compared to simulations (i.e. the bias and variance on 300 simulations were not significantly different from the expected ones).

The bias on $\hat{g}_2$ is positive and increases with the selfing rate; however, it remains very small (relative to the true value of g_2) even for large s. The bias in s (estimated either from $\hat{g}_2$ or by the ML method) is so small that it can be considered negligible for practical purposes. A very slight positive bias (~1%) is expected when $s = 0$, which is unavoidable as negative estimates are not possible. This

© 2007 Centre National de la Recherche Scientifique
Journal compilation © 2007 Blackwell Publishing Ltd

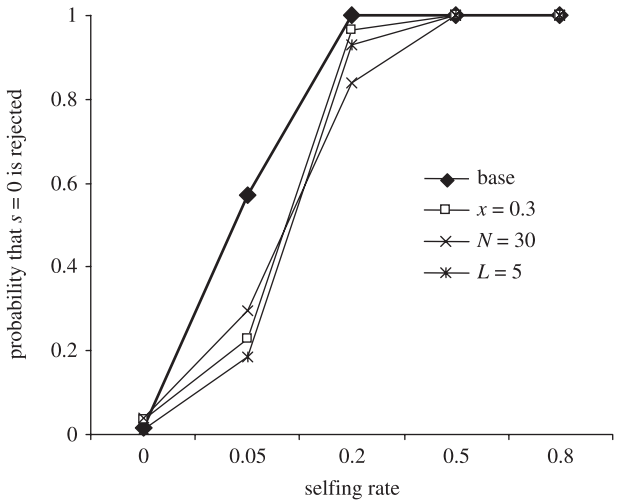

Fig. 2 The power to reject the null hypothesis $s = 0$ at $P < 0.05$, using the maximum-likelihood method. The method based on resampling g_2 values gives very similar results and has not been represented.

contrasts with the bias expected when s is estimated using F_{IS}, which is of the order of magnitude of the misscoring rate x (10–30% in our examples), or even twice the misscoring rate, when s is low. The MSE (mean squared error, combining bias and sampling variance) of $\hat{s}$ is given in Fig. 1(D), together with that expected for F_{IS}-based estimates of s. The MSE of F_{IS} is mostly driven by the bias due to misscoring, so we also represent its value without misscoring ($x = 0$) which represents pure sampling variance. The MSE of ML or $\hat{g}_2$-based estimates (even with 10% misscoring) is similar to the error variance expected for F_{IS}-based estimates without misscoring, although slightly smaller when $s = 0$ and slightly higher for intermediate s (e.g. $s = 0.5$). The MSE of F_{IS}-based estimates with misscoring ($x = 0.1$) is much higher (because of the bias), except for very high s, when the bias is constrained by the condition $s < 1$. Also, the 95% confidence intervals (CI) on s obtained by the ML method turned out to be accurate or slightly too conservative in all cases (i.e. with all the parameter sets tested, the percentage of simulations in which the CI contained the true value of s was on average 95% (minimum 93.7% maximum 99%).

The power to reject the null hypothesis $s = 0$ is indicated in Fig. 2. It is difficult here to compare with the F_{IS} method because, with misscoring rates of 0.1 or 0.3, F_{IS} is always significant, leading to a systematic rejection of the null hypothesis $s = 0$ even when it is true. Figure 2 shows that the g_2 or ML methods do not share this sensitivity to misscoring: the rejection rate is always less than or equal to the nominal type II risk (here, 5%), when $s = 0$. There is a slight difference between the ML and $\hat{g}_2$ methods however. When $s = 0$, the rejection rate of the $\hat{g}_2$ method is 0.046 (not significantly different from 0.05) while that of the ML method is lower (0.022, significantly less than 0.05). For $s = 0.05$, the powers are the same (respectively 0.55 and 0.57) in the optimal conditions (base parameter set) but tend to be slightly lower for the ML method when the number of loci, the sample size, the genetic diversity, the accuracy of genotyping, or combinations thereof, are reduced (not shown). The ML method is therefore slightly too conservative for very low selfing rates, especially when the data sets are 'imperfect'. This does not mean that estimates are too low (biased), just that the ML method finds them non-significant slightly too often. However, the power of both methods quickly rises to 80–100% for all parameter sets when s equals 0.2 or more (Fig. 2).

The estimations of s on empirical data sets are given in Table 1 and Fig. 3. Each data set comprises several subsamples (eight annual cohorts for the population of the bivalve *Spisula ovalis*; 12 localities around Lake Geneva for *Physa*

Table 1 Estimation of the selfing rate s from empirical data sets by three methods: F_{IS}, g_2 and maximum likelihood (ML). N_{pop}, number of populations in the data set. N_{tot}, total sample size summed over populations. L, number of loci. For the g_2 and F_{IS} methods, the average parameter (overall populations), estimated selfing rate s and P value for the test of H_0: $s = 0$ are given. This P value has been obtained by combining P values of all populations using Fisher's method. For the ML method, estimates of s and 95% confidence intervals are given. The hypothesis H_0: $s = 0$ is tested by model simplification, comparing the difference in deviance to a χ^2 with 1 d.f. The last two columns refer to the test of homogeneity of s across populations (difference in deviance tested as a χ^2 with N_{pop}-1 d.f.)

				F_{IS} method			g_2 method			ML method				Homogeneity	
Species	N_{pop}	N_{tot}	L	F_{IS}	$\hat{s}(F_{IS})$	$P_{F_{IS}=0}$	Mean $\hat{g}_2$	$\hat{s}$ ($\hat{g}_2$)	$P_{g_2=0}$	$\hat{s}$ (ML)	CI95	$\delta dev_{s=0}$	$P_{s=0}$	δdev	P
Spisula ovalis	8	2812	9	0.037	0.071	**0.00**	0.003	0.012	0.124	0.016	[0, 0.042]	1.680	0.195	4.95	0.67
Physa acuta	12	460	6	0.297	0.458	**0.00**	0.008	0.031	0.275	0.000	[0, 0.083]	0.000	1.000	5.11	0.93
Manihot esculenta	2	395	8	0.071	0.132	**0.00**	0.057	0.182	**0.000**	0.163	[0.090, 0.231]	19.88	**0.000**	0.22	0.64
Bulinus truncatus	3	177	4	0.828	0.906	**0.00**	1.873	0.855	**0.000**	0.841	[0.816, 0.875]	66.60	**0.000**	0.41	0.814

Significant P-values are in bold.

© 2007 Centre National de la Recherche Scientifique
Journal compilation © 2007 Blackwell Publishing Ltd

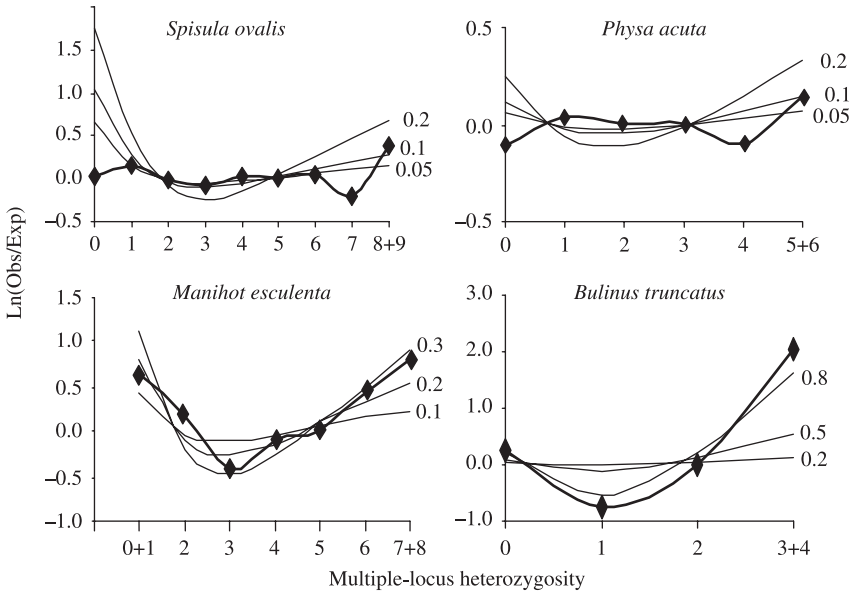

Fig. 3 Plots of the log-ratio of observed over expected number of individuals in each heterozygosity class, for the four data sets analysed. Expected numbers are obtained under independent assortment of single-locus heterozygosities (i.e. under random mating). Theoretical graphs (thin lines) for different selfing rates are given together with empirical data (thick lines). Extreme MLH classes have been pooled together to get more than five expected individuals within each class.

acuta, two fields for *Manihot esculenta*, and three populations for *Bulinus truncatus*). In all cases, the likelihood-ratio test indicates no significant difference in selfing rates among subsamples, so we present pooled statistics in Table 1 (the population details are in Appendix S2, Supplementary material). The data on *S. ovalis* and *P. acuta* are consistent with $s = 0$ using both the $\hat{g}_2$ and ML methods, although the F_{IS} are very significant and yield substantial selfing rates of 7% and 45%, respectively. On the other hand, the three methods give comparable estimates for *M. esculenta* (a moderate selfer) and *B. truncatus* (a highly selfing species), although the F_{IS}-based estimates are slightly above the other two in the latter. Figure 3 allows visualizing heterozygosity disequilibria. The two species in which the $\hat{g}_2$ and ML methods reveal no inbreeding (*S. ovalis* and *P. acuta*) essentially show a good agreement with frequencies predicted by random assortment of single-locus genotypes (flat graph). On the other hand, *M. esculenta* and *B. truncatus* yield U-shaped graphs, in good agreement with expectations based on their estimated selfing rate (see Fig. 3). The figure also illustrates an important characteristic of this method, i.e. the important role of extreme heterozygosity classes (very homozygous and very heterozygous individuals) to infer selfing rates. For example, in *B. truncatus*, most individuals originate from several generations of selfing, and are homozygous at all four loci, but a few of them are outcrossed, and are often heterozygous at three or four loci at the same time (a very unlikely event if loci were statistically independent). In this case, such individuals are quite easy to spot in the data set. For moderate selfers such as *M. esculenta*, the principle is the same although it is impossible to tell apart selfed and outcrossed individuals precisely.

Discussion

The basic rationale behind this work is that genotypes contain two different kinds of information about selfing, both of which are correlations among genes within individuals: within-locus correlations (i.e. excess homozygosity) and among-locus correlations (i.e. identity disequilibria). The different methods available to estimate selfing make use of either one or both sources of information. F_{IS}-based estimates, the Bayesian and maximum-likelihood single-locus population estimates available (Hill *et al*. 1995; Ayres & Balding 1997), as well as single-locus estimates from progeny arrays (Brown & Allard 1970) deal only with within-locus correlations. Multilocus estimates from progeny arrays (Ritland & Jain 1981; Shaw *et al*. 1981) and the maximum-likelihood method of Enjalbert & David (2000) use both types of correlations. The present methods use only among-locus correlations. In principle, one might expect that the most precise estimates will be obtained using a maximum of information in the data, i.e. by combining within- and between-locus correlations. However, in practice, two limits exist. First, the number of 'nuisance' parameters to be estimated increases in proportion to the complexity taken into account. For example, one can generalize the maximum-likelihood approach to use multilocus genotypes, including not only heterozygosity (as here) but also allelic states. However, this imposes the joint estimation of a huge number of parameters (allelic frequencies for all alleles in each locus, and misscoring rates for all allele pairs) together with the parameter of interest. One may avoid this by setting allele frequencies to constants and ignoring sampling error on them. However, this will discard the main advantage of the maximum-likelihood

© 2007 Centre National de la Recherche Scientifique
Journal compilation © 2007 Blackwell Publishing Ltd

approach, which is to be based on an explicit, general model. As soon as one enters sample-dependent constants in the model, it is not general anymore. Allele frequencies could, of course, be estimated on a separate sample much larger than the one used for estimating selfing rates; however, this is tedious and, in practice, rarely done. A second limit, maybe more important in practical terms, is that not all sources of information are equally reliable. In that case, it is a good idea to compare estimates from independent sources rather than using only one of them, or collapsing them together. Within-locus correlations are often less reliable than among-locus correlations because the main sources of misscoring (null alleles and partial dominance) only affect within-locus correlations. The methods presented here allow extracting selfing rates exclusively from among-locus correlations in multilocus population data. Ignoring allele frequencies seems to decrease precision, for example a homozygote for a rare allele is very likely an inbred and this information will be ignored by our method. However, in the presence of artefacts, such genotypes also have the largest chance of being spurious. To avoid such risks, our method should be used in addition to the F_{IS}-method (within-locus correlations) or other allele-frequency methods, to get a realistic interpretation. In case of disagreement, especially if s appears larger using F_{IS}, technical artefacts should be suspected.

Are technical biases frequent enough to worry about? The answer is probably yes (Bonin *et al*. 2004). Significant heterozygote deficiencies are the rule rather than the exception in both allozyme and microsatellite data, even for dioecious or self-incompatible organisms unable to self (Zouros & Foltz 1984; Bonin *et al*. 2004). Of course, nontechnical causes such as population substructure or homogamy can generate heterozygote deficiencies. However, F_{IS} values are often higher than F_{ST} among neighbouring populations or patches, making substructure an unlikely explanation (David *et al*. 1997b); similarly, homogamy is quite difficult to imagine for molecular markers with little or no phenotypic effect. Therefore, inbreeding rates inferred from F_{IS} should be treated very cautiously, especially in predominantly outbred species in which the relative error can be large. The progeny-array method is expected to be less sensitive to technical biases because such biases can generate inconsistent genotypes (i.e. offspring with no allele from their mother) that can at least be noticed and raise suspicion. However, we are aware of no study of the robustness of progeny-array estimates with respect to technical biases.

Our simulations show that methods based on the among-locus correlation structure of the data ($\hat{g}_2$ and ML) have useful statistical properties. With realistic sample sizes and genetic diversity, their statistical bias is negligible, irrespective of misscorings, and of the true value of s. Their precision (SE) and, subsequently, the power to reject $s = 0$ are very slightly lower than that of F_{IS} (averaged over the same number of loci) when s is intermediate or high, but higher when s is small (Fig. 1). Whatever the method, a useful starting point to evaluate the sample size necessary to achieve a given precision in s, is to consider binomial variance, i.e. the variance that would remain if we knew for each individual its selfed or outcrossed status without error. For example, for $N = 30$ and $s = 0.1$, the binomial 95% CI on the proportion of selfed individuals in the sample is [0, 0.23], and no statistical method, however, precise, will give much better results. The two methods (based on $\hat{g}_2$ and ML) have very comparable performances in terms of power and accuracy, although the ML method is slightly too conservative for very small s. ML has the advantage of providing simple hypothesis tests on the value of s or on its variation among samples. However, checking that the value of $\hat{g}_2$ and the form of the MLH distribution (see Fig. 3) are consistent with ML results is always a good idea.

We have illustrated the behaviour of the two methods with four real data sets. *Spisula ovalis* is a dioecious species. Yet, previous studies have suggested that some other form of inbreeding occurs (at low rates) and generates heterozygosity-fitness correlations in this species (David *et al*. 1997a). The significant increase in heterozygosity with age (David & Jarne 1997) suggests that inbred individuals are progressively removed by selection. We therefore expected young cohorts (rather than old ones) to show traces of inbreeding, in the form of positive, if low, estimates of selfing rates. The F_{IS} method provides significantly positive estimates for all cohorts (Table 1 and Appendix S2, Supplementary material) but the detailed study of the variation of F_{IS} among loci, sites, and cohorts in that sample has led to the conclusion that null alleles, rather than any population phenomenon, were likely to account for most of them (David *et al*. 1997b). The multilocus methods ($\hat{g}_2$ and ML) both give nonsignificant estimates, much lower than F_{IS}-based estimates. This is not contradictory with the expected low, but positive, biparental inbreeding initially expected. Although the power is too low to reach significance in the pooled sample, there is a trend of decreasing s from young to old cohorts (as expected, see above), and the s is marginally significant in the youngest cohort (1 year) using the less conservative of the two methods ($\hat{g}_2$). At any rate, the confidence intervals on s largely allow for a rate of inbreeding equivalent to a few per cent of selfing in the young cohorts. The discrepancy between F_{IS}-based and multilocus-based estimates is consistent with a low misscoring rate ($x = 2.9\%$, averaged over loci). Null alleles with frequency 1.5% or so would be sufficient to generate this discrepancy, and yet would remain undetected as the null homozygotes would be vanishingly rare.

The case of *Physa acuta* is, to an even greater extent than *S. ovalis*, illustrative of the unreliability of F_{IS}-based estimates. Multilocus methods identify no hint of self-fertilization in this hermaphroditic species, in agreement with the selfing

© 2007 Centre National de la Recherche Scientifique
Journal compilation © 2007 Blackwell Publishing Ltd

rates of 0–0.1 obtained using allozyme or microsatellite-based progeny arrays in other populations (Jarne *et al*. 2000; Henry *et al*. 2005). F_{IS}-based estimates suggest large selfing rates (close to 50%) and are probably artefactual. This is certainly suggested by the wide range of variation of F_{IS}-estimates among loci within populations (from −0.16 to 0.71). The low polymorphism at many loci may decrease the power to detect multilocus associations within each population; however, in the combined data set, the sample size is quite large and the estimates still do not come close to significance. The confidence intervals clearly exclude high values of s in this case, and the implied misscoring rate is close to 30%.

In the *Manihot* data set, in contrast, the multilocus structure shows the expected U shape for a moderately selfing population and the multilocus estimates (by either method) give significant values in the range 0.15–0.2. Note that this is not necessarily 'selfing' in its usual sense. Traditional cassava farmers in French Guiana (where the two fields were sampled) propagate plants clonally and distinguish several clonal varieties, purposely planting a few of them in distinct patches of their fields (Pujol *et al*. 2005). 'Selfing' here may account for pollination events between clonemates rather than by the mother plant. However, neglecting mutational variation arising during clonal propagation, the consequences are the same.

Finally, the case of *Bulinus truncatus* illustrates the behaviour of the multilocus estimates with very high selfing rates (0.8–0.9). One can expect our methods to lose power (unlike F_{IS}-based methods) in this case because very few heterozygotes are found in such populations. Therefore, these methods are bound to fail when $s = 1$, because they rely on the clustering of heterozygosity within individuals. However, the example of *Bulinus*, as well as our simulations, show that up to very large selfing rates, this clustering remains detectable and allows a surprisingly precise estimation of s, in agreement (in this case) with F_{IS}-based estimates and progeny-array analyses (Viard *et al*. 1997a). This is due to the fact that outcrossing events, as few as they might be, are readily detectable (multiple heterozygotes, very unlikely under independence), and that the $\hat{g}_2$ and ML methods base their estimation on the whole pedigree, i.e. incorporate not only the difference between selfed and outcrossed, but also between one and several generations of selfing, as in Enjalbert & David 2000). When s is high enough, this allows multilocus estimates to have a variance even lower than binomial, i.e. lower than it would be if only two classes of individuals (selfed and outcrossed) could be recognized without error (data not shown).

The final question that remains concerns the population model underlying our estimation, i.e. an unstructured mixed-mating population at linkage and inbreeding equilibrium. No estimate can be accurate if the underlying model is a bad description of reality. It is quite clear that very few populations will fulfil all the assumptions, the problem is rather to know to what extent realistic deviations from the model can affect the estimates and their biological significance. Inbreeding equilibrium is a parsimonious assumption unless one suspects recent changes in selfing rates (see Enjalbert & David 2000 for a way to detect such historical changes from multilocus genotypes). Because of the geometric distribution of the number of selfing generations, and of the resetting effect of outcrossing, inbreeding nonequilibrium can be a serious problem only for high s (say > 0.7). This bias is shared by F_{IS}-based estimates.

A more complex problem lies with the fact that multilocus zygotic associations, or identity disequilibria, which constitute the basis of our estimates, can be affected by other phenomena than selfing. Gametic disequilibria within samples, for example due to population substructure, are known to produce multilocus zygotic associations (Yang 2000). The excess of multihomozygotes or multiheterozygotes, either in real genotypes, or in theoretical genotypes generated by random re-association of real haplotypes, has even been used as a measure of gametic disequilibrium (Brown *et al*. 1980; Sabatti & Risch 2002) in populations assumed to be at HW equilibrium. Our methods could therefore inaccurately estimate positive selfing rates in such populations. The only solution to this problem is to carefully evaluate the possibility of gametic disequilibria. In many situations, gametic disequilibria are likely to be small, because the population is large, the spatial scale of sampling is small relative to dispersion, and the markers are on different chromosomes. This can be checked using tests of gametic disequilibria (e.g. based on R^2 or D estimates). Note that although gametic disequilibria usually tend to increase the variance in heterozygosity (and thus $\hat{g}_2$ and $\hat{s}$), this is not necessarily the case, depending on the exact configurations. Overall, the configurations that most affect the variance are preferential associations between alleles of similar frequencies at different loci (increase in variance) or between alleles of very different frequencies (decrease in variance) (Yang 2000, 2002). Averaged over many pairs of loci and alleles (usually physically unlinked), neither case is very likely to be frequent in standard population genetic studies. An interesting case is linkage disequilibrium due to population substructure (multilocus Wahlund effect). Mixing different subpopulations within the sample will affect our estimates of selfing only to the extent that it changes the MLH distribution. This happens when subpopulations with different average heterozygosities (over all loci) are mixed together, which results in a mix of high-heterozygosity and low-heterozygosity genotypes and therefore mimics the effects of selfing. An advantage of our multilocus method compared to F_{IS} is that the latter is sensitive to any heterogeneity in allele frequencies between subpopulations, whereas our multilocus method is sensitive to a much more restricted set of differences: those that make one subpopulation systematically less heterozygous

© 2007 Centre National de la Recherche Scientifique
Journal compilation © 2007 Blackwell Publishing Ltd

(on average over all loci) than the other. Unless populations with very different effective sizes are mixed together, the latter case is probably not frequent.

However, there are definitely some special cases, such as when hybridization between two distinct taxa is suspected (Barton & Hewitt 1985), or when loci are extremely linked, for example different nucleotidic sites within a sequence (e.g. Sabatti & Risch 2002), where gametic disequilibria are likely, and in such cases, our estimates of selfing rate should therefore not be used. At the other extreme, populations with high rates of self-fertilization often have low effective sizes (Doums *et al*. 1996; Charlesworth 2003), which together with their low effective recombination rates (even for physically unlinked loci) can generate large gametic disequilibria, as the population becomes more or less structured into homozygous selfing lines. In this case, there may be a risk of overestimating s, as crosses between two individuals within such lines mimic true selfing. This problem usually remains minor because s has to be very high (and N_e small) to generate such a structure; and then the overestimation is small because of the maximum value $s = 1$. In most data sets (N of the order of 100), the possible bias due to gametic disequilibria in highly selfing populations will be small compared to the binomial sampling error.

Apart from linkage disequilibria, biparental inbreeding (mating between relatives) and selection may affect the estimates of s. Inbreeding depression, a widespread form of selection (Charlesworth & Charlesworth 1987), is expected to decrease the apparent selfing rate as a function of the age of individuals, leading to underestimation of 'primary' selfing rates (Lande *et al*. 1994). We note, however, that biases due to gametic disequilibria, biparental inbreeding, and selection are shared by all methods to estimate selfing rates published so far. The only exception is that progeny-array analysis allows the estimation of biparental inbreeding in addition to selfing rates (Ritland 2002). Inbreeding depression is usually considered the most likely source of systematic bias in estimates of selfing rates; however, because inbreeding depression progressively removes inbred individuals as the cohort grows older, none of the methods can by itself avoid this bias. The only way out is to sample as young individuals as possible to come closer to the primary inbreeding or selfing rate, as exemplified by the *Spisula* data set, in which inbreeding is detected in young, but not old, cohorts. In comparison to progeny-array approaches, methods based on population data (F_{IS} and multilocus methods presented in this paper) are expected to be even more sensitive to inbreeding depression, because the inbreeding equilibrium assumption is violated. However this additional effect is likely to be small in comparison to the direct effect of removing recent inbreds, because inbreeding depression is likely to be strong in predominantly outbred species, where inbreeding equilibrium is not a crucial issue (see above).

Finally, in many situations, the estimate of selfing rate has a simple and biologically meaningful interpretation even if affected by other phenomena. This is the case, for example, in two of our data sets. In *S. ovalis*, a dioecious species, s rather represents some form of biparental inbreeding than selfing; and its decrease in successive cohorts can be interpreted as an effect of inbreeding depression (David *et al*. 1997a). In the *Manihot* data set, as already stated, matings between clonemates planted in patches within each field is likely to account for apparent selfing. Strictly speaking, this is a case of extreme population substructure and linkage disequilibrium, rather than selfing. However the predicted consequences, i.e. the occurrence of inbreeding depression and relationships between heterozygosity and fitness traits, are the same, and are readily observed (Pujol *et al*. 2005; Pujol & McKey 2006). Symmetrically, true selfing can be viewed just as an extreme case of population substructure with subpopulations of size $N = 1$.

In conclusion, our multilocus methods to estimate s may sometimes confound selfing with some other population phenomena, just like other methods do — and little more than biological knowledge of the system and good sense can fix this. However, unlike others, it does not confound selfing with (presumably ubiquitous) genotyping artefacts. This is a crucial issue, especially for the study of animal mating systems. One of the main contemporary issues in animal mating system evolution, i.e. the very existence of mixed-mating strategies (s close to 0.5) in animal species, is at stake. Indeed, although many species were classified as mixed-maters based on F_{IS} (Jarne & Auld 2006), these species may well turn out to be downright outcrossers once typing artefacts are removed, as illustrated by our *Physa* data set. We believe that, having to live in the real world, with null alleles, ambiguous band patterns, and human error (Pompanon *et al*. 2005), population geneticists might often find our methods useful to avoid misinterpretation of data.

Acknowledgements

We thank P. Jarne, M.-F. Ostrowski, and the GENDYN team (CEFE) for helpful comments on this manuscript. This work has been funded by grants from the CNRS and French Ministry of Environment.

References

Ayres KL, Balding DJ (1997) Measuring departures from Hardy–Weinberg: a Markov chain Monte-Carlo method for estimating the inbreeding coefficient. *Heredity*, **80**, 769–777.

Barton NH, Hewitt GM (1985) Analysis of hybrid zones. *Annual Review of Ecology and Systematics*, **16**, 113–148.

Belkhir K, Borsa P, Chikhi L, Raufaste N, Bonhomme F (2004) *GENETIX 4.05, Logiciel sous Windows pour la Génétique des Populations*. Laboratoire Génome, Populations, Interactions, CNRS UMR 5171, Université de Montpellier, France.

Bennett JH, Binet FE (1956) Association between Mendelian factors with mixed selfing and random mating. *Heredity*, **10**, 51–56.

© 2007 Centre National de la Recherche Scientifique
Journal compilation © 2007 Blackwell Publishing Ltd

Bierne N, Tsitrone A, David P (2000) An inbreeding model of associative overdominance during a population bottleneck. *Genetics*, **155**, 1981–1990.

Björklund M (2005) A method for adjusting allele frequencies in the case of microsatellite allele drop-out. *Molecular Ecology*, **5**, 676–679.

Bonin A, Bellemain E, Bronken Eidesen P *et al*. (2004) How to track and assess genotyping errors in population genetics studies. *Molecular Ecology*, **13**, 3261–3273.

Brookfield JFY (1996) A simple method for estimating null allele frequency from heterozygote deficiency. *Molecular Ecology*, **5**, 453–455.

Brown AHG, Allard RW (1970) Estimation of the mating system in open-pollinated maize populations using isozyme polymorphisms. *Genetics*, **66**, 133–145.

Brown AHD, Feldman MW, Nevo E (1980) Multilocus structure of natural populations of *Hordeum spontaneum*. *Genetics*, **96**, 523–536.

Charlesworth D (2003) Effects of inbreeding on the genetic diversity of populations. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **358**, 1051–1070.

Charlesworth D, Charlesworth B (1987) Inbreeding depression and its evolutionary consequences. *Annual Review of Ecology and Systematics*, **18**, 237–268.

David P, Delay B, Jarne P (1997a) Heterozygosity and growth in the marine bivalve *Spisula ovalis*: testing alternative hypotheses. *Genetical Research*, **70**, 215–223.

David P, Jarne P (1997) Context-dependent survival differences among electrophoretic genotypes in natural populations of the marine bivalve *Spisula ovalis*. *Genetics*, **146**, 335–344.

David P, Perdieu MA, Pernot AF, Jarne P (1997b) Fine-grained spatial and temporal genetic structure in the marine bivalve *Spisula ovalis*. *Evolution*, **51**, 1318–1322.

Doums C, Viard F, Pernot AF, 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.

Fyfe JL, Bailey NTJ (1951) Plant breeding studies in leguminous forage crops. I. Natural cross-breeding in winter beans. *Journal of Agricultural Science*, **41**, 371–378.

Gaffney PM, Scott TM, Koehn RK, Diehl WJ (1990) Interrelationships of heterozygosity, growth rate and heterozygote deficiencies in the Coot Clam, *Mulinia lateralis*. *Genetics*, **124**, 687–699.

Goodwillie C, Kalisz S, Eckert CG (2005) The evolutionary enigma of mixed mating systems in plants: Occurrence, theoretical explanations and Empirical evidence. *Annual Review of Ecology and Systematics*, **36**, 47–79.

Henry PY, Bousset L, Sourrouille P, Jarne P (2005) Partial selfing, ecological disturbance and reproductive assurance in an invasive freshwater snail. *Heredity*, **95**, 428–436.

Hill WG, Babiker KA, Ranford-Cartwrigh LC, Walliker D (1995) Estimation of inbreeding coefficients from genotype data on multiple alleles, and application to estimation of clonality in malaria parasites. *Genetical Research*, **65**, 53–61.

Hoffman JI, Amos W (2005) Microsatellite genotyping errors: detection approaches, common sources and consequences for paternal exclusion. *Molecular Ecology*, **14**, 599–612.

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. *Annual Review of Ecology and Systematics*, **24**, 441–466.

Jarne P, Perdieu MA, Pernot AF, Delay B, David P (2000) The influence of self-fertilization and grouping on fitness attributes in the freshwater snail *Physa acuta*: population and individual inbreeding depression. *Journal of Evolutionary Biology*, **13**, 645–655.

Lande R, Schemske DW, Schultz ST (1994) High inbreeding depression, selective interference among loci, and the threshold selfing rate for purging recessive lethal mutations. *Evolution*, **48**, 964–978.

McCullagh P, Nelder JA (1989) *Generalized Linear Models*. Chapman & Hall, London.

Pasteur N, Pasteur G, Bonhomme F, Catalan J, Britton-Davidian J (1987) *Manuel Technique de Génétique Par Électrophorèse Des Protéines*. Lavoisier, Paris.

Pompanon F, Bonin A, Bellemain E, Taberlet P (2005) Genotyping errors: causes, consequences and solutions. *Nature Reviews Genetics*, **6**, 847–859.

Pujol B, David P, McKey D (2005) Microevolution in agricultural environments: how a traditional Amerindian farming practice favours heterozygosity in cassava (*Manihot esculenta* Crantz, Euphorbiaceae). *Ecology Letters*, **8**, 138–147.

Pujol B, McKey D (2006) Size asymmetry in intraspecific competition and the density-dependence of inbreeding depression in a natural plant population: a case study in cassava (*Manihot esculenta* Crantz, Euphorbiaceae). *Journal of Evolutionary Biology*, **19**, 85–96.

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

Ritland K, Jain SK (1981) A model for the estimation of outcrossing rate using *n* independent loci. *Heredity*, **47**, 35–52.

Robertson A, Hill WG (1984) Deviations from Hardy–Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. *Genetics*, **107**, 703–718.

Sabatti C, Risch N (2002) Homozygosity and linkage disequilibrium. *Genetics*, **160**, 1707–1719.

Schemske DW, Lande R (1985) The evolution of self-fertilization and inbreeding depression in plants. II. Empirical observations. *Evolution*, **39**, 41–52.

Shaw DV, Kahler AL, Allard RW (1981) A multilocus estimator of mating system parameters in plant populations. *Proceedings of the National Academy of Sciences, USA*, **78**, 1298–1302.

Van Oosterhout C, Hutchinson WF, Wills DPM, Shipley P (2004) MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data. *Molecular Ecology*, **4**, 535–538.

Van Oosterhout C, Weetman D, Hutchinson WF (2006) Estimation and adjustment of microsatellite null alleles in nonequilibrium populations. *Molecular Ecology Notes*, **6**, 255–256.

Viard F, Doums C, Jarne P (1997a) Selfing, sexual polymorphism and microsatellites in the hermaphroditic freshwater snail *Bulinus truncatus*. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **264**, 39–44.

Viard F, Justy F, Jarne P (1997b) The influence of self-fertilization and population dynamics on the genetic structure of subdivided populations: a case study using microsatellite markers in the freshwater snail *Bulinus truncatus*. *Evolution*, **51**, 1518–1528.

Wattier R, Engel CR, Saumitou-Laprade P, Valero M (1998) Short allele dominance as a source of heterozygote deficiency at microsatellite loci: experimental evidence at the dinucleotide locus Gv1CT in *Gracillaria gracilis* (Rhodophyta). *Molecular Ecology*, **7**, 1569–1573.

Weir BS, Avery PJ, Hill WG (1980) Effect of mating structure on variation in inbreeding. *Theoretical Population Biology*, **18**, 396–429.

Weir BS, Cockerham CC (1973) Mixed self and random mating at two loci. *Genetical Research*, **21**, 247–262.

© 2007 Centre National de la Recherche Scientifique
Journal compilation © 2007 Blackwell Publishing Ltd

Weir BS, Cockerham CC (1984) Estimating *F*-statistics for analysis of population structure. *Evolution*, **38**, 1358–1370.
Yang RC (2000) Zygotic associations and multilocus statistics in a nonequilibrium diploid population. *Genetics*, **155**, 1449–1458.
Yang RC (2002) Analysis of multilocus zygotic associations. *Genetics*, **161**, 435–445.
Zouros E, Foltz DW (1984) Possible explanations of heterozygote deficiency in bivalve molluscs. *Malacologia*, **25**, 583–591.

All authors are population geneticists interested in the evolution of mating systems. They focus on empirical work using different model systems: marine and freshwater invertebrates, especially snails (P. David, F. Viard, V. Castella, J. Goudet); flowering plants (B. Pujol, J. Goudet) as well as theoretical issues (P. David, J. Goudet).

Supplementary material

The following supplementary material is available for this article:

Appendix S1 Statistical properties of the estimator of second-order heterozygosity disequilibria

Appendix S2 Estimation of selfing rates by population

This material is available as part of the online article from:
http://www.blackwell-synergy.com/doi/abs/10.1111/j.1365-294X.2007.03330.x
(This link will take you to the article abstract).

Please note: Blackwell Publishing are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

© 2007 Centre National de la Recherche Scientifique
Journal compilation © 2007 Blackwell Publishing Ltd

Appendix

The effect of misscoring at one locus on the estimation of F_{IS}

The effect of misscoring on the estimation of F_{IS} will obviously depend on which kind of artefacts affect the apparent heterozygosity and on which estimator is chosen. To simplify the picture, we will consider a single bi-allelic locus, and the simple estimator of F_{IS} based on heterozygote deficiency (f_1). This can be generalized to several alleles using estimators designed for this case (Robertson & Hill 1984; Weir & Cockerham 1984). The heterozygote deficiency is:

$$\hat{f}_1 = 1 - \frac{\sum H_j/N}{\hat{D}} \qquad \text{(eqn A1)}$$

where the summation is over individuals $j = 1 \ldots n$ (no subscript for locus here as we consider only one locus), and $\hat{D} = [(N-1)/N]2\hat{p}(1-\hat{p})$ is the estimated gene diversity, given $\hat{p}$ the estimated allele frequency in the sample. The factor $(N-1)/N$ corrects for finite sample size. The artefacts responsible for misscoring of heterozygotes (e.g. null alleles) can simultaneously bias the estimation but the amount of bias (say y) depends on the exact kind of artefact considered. Hence the knowledge of x is not sufficient to obtain y unless we assume some particular kind of misscoring.

Accounting for a rate x of misscoring of heterozygotes, and a bias y in the estimation of D (note that y may be zero), and noting X the parameter defined by $(1-X) = (1-x)/(1-y)$ the expected value of $\hat{f}_1$ is approximately:

$$E(\hat{f}_1) = 1 - (1-X)(1-E(f)) = E(f) + X(1-E(f)) \qquad \text{(eqn A2)}$$

in which $X(1-E(f))$ represents the absolute bias in the estimate of $E(f)$. This estimation neglects small-order terms arising from the difference between the expectation of a ratio and the ratio of expectations; Weir & Cockerham (1984) have shown that this usually has little consequence. Note also that in many cases (e.g. allele dropout) y is expected to be relatively small and then x is a good approximation of X.

© 2007 Centre National de la Recherche Scientifique
Journal compilation © 2007 Blackwell Publishing Ltd