2011

Inbreeding effective population size and parentage analysis without parents

Robin S. Waples
NOAA Fisheries, robin.waples@noaa.gov

Ryan K. Waples
Casa Azul

MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2011) **11 (Suppl. 1),** 162–171 doi: 10.1111/j.1755-0998.2010.02942.x

ANALYTICAL APPROACHES

Inbreeding effective population size and parentage analysis without parents

ROBIN S. WAPLES* and RYAN K. WAPLES†
*NOAA Fisheries, Northwest Fisheries Science Center, 2725 Montlake Blvd. East, Seattle, WA 98112, USA, †Casa Azul, 10056 Dibble Ave N, Seattle, WA 98177, USA

Abstract

An important use of genetic parentage analysis is the ability to directly calculate the number of offspring produced by each parent (k_i) and hence effective population size, N_e. But what if parental genotypes are not available? In theory, given enough markers, it should be possible to reconstruct parental genotypes based entirely on a sample of progeny, and if so the vector of parental k_i values. However, this would provide information only about parents that actually contributed offspring to the sample. How would ignoring the 'null' parents (those that produced no offspring) affect an estimate of N_e? The surprising answer is that null parents have no effect at all. We show that: (i) The standard formula for inbreeding N_e can be rewritten so that it is a function only of sample size and $\sum(k_i^2)$; it is not necessary to know the total number of parents (N). This same relationship does not hold for variance N_e. (ii) This novel formula provides an unbiased estimate of N_e even if only a subset of progeny is available, provided the parental contributions are accurately determined, in which case precision is also high compared to other single-sample estimators of N_e. (iii) It is not necessary to actually reconstruct parental genotypes; from a matrix of pairwise relationships (as can be estimated by some current software programs), it is possible to construct the vector of k_i values and estimate N_e. The new method based on parentage analysis without parents (PwoP) can potentially be useful as a single-sample estimator of contemporary N_e, provided that either (i) relationships can be accurately determined, or (ii) $\sum(k_i^2)$ can be estimated directly.

Keywords: contemporary N_e, relationship, reproductive success, sib reconstruction, single-sample estimator

Received 25 May 2010; revision received 12 August 2010; accepted 18 August 2010

Introduction

In the last two decades, parentage analyses made possible by highly polymorphic molecular markers have produced many novel insights (reviewed by Jones *et al.* 2010). In a typical parentage analysis, multilocus genotypes are scored in both progeny and potential parents, and these data are used to 'assign' progeny to parents, either through a probabilistic framework or by excluding other potential parents as impossible, given the rules of Mendelian inheritance. Parentage analysis can provide diverse types of information, including the direct (demographic) calculation of effective population size (N_e), which is one of the most important parameters in evolutionary biology but also one of the most difficult to estimate. For a monoecious population with random selfing, inbreeding effective size is given by (Crow & Denniston 1988, Equation 1; Caballero 1994):

Correspondence: Robin Waples, Fax: (206) 860 3335;
E-mail: robin.waples@noaa.gov

$$N_e = \frac{\bar{k}N - 1}{\bar{k} - 1 + V_k/\bar{k}} \qquad \text{(eqn 1)}$$

where N is the number in the parental generation, and $\bar{k}$ and V_k are the mean and variance of the number of offspring contributed by each parent (k_i). If genetic methods can allow one to identify the number of offspring produced by each parent, the vector of k_i values can be used to calculate $\bar{k}$, V_k and N_e using eqn (1) (or slight variations that apply to other mating systems).

Now consider a variation to this scenario in which the parents cannot be sampled, but parents can be reconstructed from genotypes of the progeny—that is, a parentage analysis without parents (PwoP). It does not appear that a complete parentage analysis without parents has been conducted to date, but under some conditions, it is possible to reconstruct genotypes of some missing parents (e.g., Emery *et al.* 2001; Wang 2004). With enough highly variable markers, this can be quite feasible, provided that a sufficient number of known siblings are available (Jones & Avise 1997; DeWoody *et al.* 2000) or at least one of the parents can be genotyped (Myers &

Zamudio 2004; Jones 2005; Tatarenkov *et al*. 2008; Eriksson *et al*. 2010). For the moment, assume that it is possible, from a sample of progeny alone, to reconstruct all genotypes from contributing parents, and from this, the vector of k_i values for those parents. But there is an important difference between this vector and the vector of k_i values in a standard parentage analysis: Because the PwoP vector only provides information about parents that actually produced progeny, it contains no elements with $k_i = 0$, which represent null parents (those that produced no progeny that survived to the stage at which they were enumerated). In a typical population, the parental generation might have included an unknown (but potentially large) number of individuals that produced no surviving offspring. For example, in an ideal population (in which $V_k \approx \bar{k} = 2$), on average, about 13% of parents will produce no offspring that survive to maturity, and the proportion should be higher in most natural populations (in which V_k typically will be larger than $\bar{k}$). What effect do these missing parents have on calculations of N_e based on PwoP? The surprising answer is that they have no effect at all: *it is not necessary to know anything about null parents to compute* N_e. We demonstrate this with a simple analytical proof, which results in a novel formula for inbreeding effective size that does not depend on N. Further, we use analytical and numerical methods to demonstrate the following:

1 Even an incomplete (but random) sample of progeny provides an unbiased estimate of inbreeding N_e using the novel formula;
2 The same insensitivity to null parents does not apply to variance effective size, which reflects the number in the progeny generation and hence varies with the number of progeny sampled;
3 It is not necessary to actually reconstruct parental genotypes to calculate N_e using PwoP; from a matrix of pairwise relationships (as can be estimated by current software programs), it is possible to construct the vector of k_i values and estimate N_e;
4 Assuming sibling relationships can be accurately determined, precision of the new method based on parentage analysis without parents appears to compare favourably with single-sample estimators of N_e currently in use. However, accurately determining relationship categories is a very challenging problem.

Effective population size and null parents

Assume a population has N potential parents, and one wants to calculate effective population size using eqn (1). Further, assume that we have sampled the entire population of offspring and that the two parents of each offspring can be identified. It is then straightforward to construct the vector of k_i values that describes the number of offspring produced by each parent. Parametric formulas for mean and variance of k can be expressed as follows:

$$\bar{k} = \frac{\sum k_i}{N}$$

$$V_k = \frac{\sum \left(k_i^2\right)}{N} - \left[\frac{\sum k_i}{N}\right]^2$$

Because we are interested in effective size of a particular population in a particular generation, we follow Crow & Denniston (1988) and treat the k_i as fixed, so in computing the population variance, the $N/(N-1)$ sample-size correction is not used. Substituting the above into eqn (1) leads to

$$N_e = \frac{\frac{\sum k_i}{N} N - 1}{\frac{\sum k_i}{N} - 1 + \frac{\frac{\sum \left(k_i^2\right)}{N} - \left[\frac{\sum k_i}{N}\right]^2}{\frac{\sum k_i}{N}}}$$

which, after simplifying, yields

$$N_e = \frac{\sum k_i - 1}{\frac{\sum \left(k_i^2\right)}{\sum k_i} - 1} \qquad \text{(eqn 2a)}$$

As each diploid progeny has two parents, $\sum k_i = 2S$, where S is the number of progeny, and eqn (2a) can also be written as

$$N_e = \frac{2S - 1}{\frac{\sum \left(k_i^2\right)}{2S} - 1} \qquad \text{(eqn 2b)}$$

The surprising consequence of eqn (2) is that inbreeding N_e does not depend on parental population size (N)—only on the two summation terms $\sum k_i$ and $\sum \left(k_i^2\right)$. Further, as $\sum k_i$ is determined by the sample size, the new formula shows that inbreeding N_e depends on a single unknown term ($\sum \left(k_i^2\right)$). Because this term is not affected by any $k_i = 0$, it follows that parents that contribute no offspring have no effect on inbreeding N_e. This means that N_e can be computed directly from the progeny generation based only on information about parents that actually leave offspring, without knowing what parental N was. The independence of inbreeding N_e with N is a curious result, given that whether one includes null parents or not affects N, $\bar{k}$ and V_k. However, these parameters change in a correlated way such that the overall effect creates no change in N_e. To illustrate, consider a diploid population with $N = 10$ adults, for which the vector of k_i values is [5, 1, 2, 4, 2, 0, 2, 1, 0, 3]. For these data, $\bar{k} = 2.0$, $V_k = 2.40$ and $N_e = 8.64$ from eqn (1). Note that two of the parents produced no offspring. If we ignore

those null parents, then $N = 8$, $\bar{k} = 2.5$, $V_k = 1.75$–all different compared to the scenario that considered all 10 potential parents. However, the resulting N_e for the scenario that ignores null parents is 8.64—identical to the value based on all the parents.

It is easy to show that this same approach does not work (at least in a general way) with variance N_e. The analogue to eqn (1) for variance N_e in a monoecious population is (Crow & Denniston 1988, Equation 19)

$$N_e = \frac{4N' - \bar{k}}{2\left[1 + V_k/\bar{k}\right]} \qquad \text{(eqn 3)}$$

where N' is the number of progeny rather than the number of parents (N). If we note that $N' = N\bar{k}/2$, this reduces to a form that is expressed in terms of the number of parents:

$$N_e = \frac{\bar{k}(2N - 1)}{2\left[1 + V_k/\bar{k}\right]} \qquad \text{(eqn 4)}$$

Substituting the formulas for $\bar{k}$ and V_k above does not lead to an expression that is independent of N, as occurs with inbreeding N_e. Because variance N_e reflects the number in the progeny generation (Crow 1954; cf eqn 3), it is proportional to $\bar{k}$ and hence depends on the sample size of progeny. For this reason, calculations of variance N_e based on parent-offspring data are generally meaningful only if the mean and variance of k are scaled to what they would be expected to be in a population of constant size (Crow & Morton 1955). The effects on PwoP analyses can be seen by considering the hypothetical example described earlier. Using data for all 10 parents, N_e calculated from eqn (4) is the same as it is for inbreeding N_e (8.64), as should always be the case for random mating populations of constant size. However, when the two null parents are excluded, variance N_e rises to 11.03, reflecting the higher $\bar{k}$ for the eight parents that actually produced offspring. In contrast, the formula for inbreeding N_e (eqn 1) has an extra $\bar{k}$ term in denominator that automatically adjusts for the effects of changes in population size. As a consequence, calculation of inbreeding N_e from large samples of progeny does not require one to first rescale $\bar{k}$ and V_k (Waples 2002), and $\hat{N}_e$ based on PwoP is not affected by changes in $\bar{k}$ when null parents are ignored. In the remainder of this study, the term N_e is used to refer to inbreeding effective size.

Reconstructing parental contributions (k_i)

Reconstructing parental genotypes from a sample of progeny is a very challenging exercise that, at present, is only feasible under special circumstances. Fortunately, doing this is not a specific requirement of estimating effective size using PwoP; all that is required is to be able to construct the vector of parental contributions (the k_i values), and this can be accomplished (or at least attempted) using information provided by sibship reconstruction programmes that are currently available (Jones *et al*. 2010). Figure 1 illustrates how information on pairwise relationships (full-sib, half-sib, unrelated) can be used to infer parental contributions. In this example, each progeny in a sample of $S = 25$ individuals is represented by a letter (A–Y). Isolated letters (C, D, F …) represent individuals that are not related to any other sampled progeny; each of these progeny must therefore have two unique parents, each with $k_i = 1$. Solid rectangles connect full siblings (e.g., H + M), and dotted ovals connect pairs of half-sibs (e.g., V + W; E + J). A pair of full siblings implies two parents each with $k_i = 2$, whereas a pair of half-siblings implies one parent with $k_i = 2$ and two parents with $k_i = 1$. More complicated, interconnected patterns of relationship are of course possible, but each can be decomposed into subsets that allow one to infer the vector of parental contributions. The only ambiguous situation of which we are aware involves a trio of individuals that are reciprocal half-siblings (e.g., offspring S, T, Y in Fig. 1). This pattern can result from either of two mating patterns: (i) one parent is responsible for all three offspring, and three parents contribute one offspring each ($k_i = 3, 1, 1, 1$) or (ii) three parents are each responsible for two of the three offspring ($k_i = 2, 2, 2$). However, the latter scenario requires at least one individual to reproduce as both male and female so is only possible for hermaphrodites. Furthermore, both of these scenarios lead to $\Sigma k_i = 6$ and $\Sigma k_i^2 = 12$, so the ambiguity has no effect on

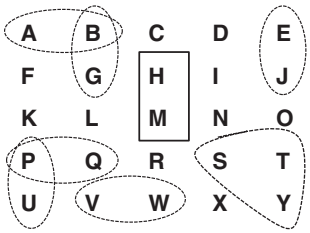

Fig. 1 Sibling relationships in a hypothetical sample of $S = 25$ progeny, each denoted by a letter. Unconnected letters (C, D, F …) represent unrelated individuals; solid rectangle includes two full siblings (H + M); dotted curves include pairs (V + W) or trios (S + T + Y) of half-siblings that all share exactly one parent. From these relationships, a vector of k_i values can be deduced. Each unrelated individual has two parents with $k_i = 1$; a pair of full sibs implies two parents, each with $k_i = 2$; a pair of half-sibs implies one parent with $k_i = 2$ and two with $k_i = 1$; and an interlocking pair of half-sib pairs (e.g., A + B + G) implies 2 parents with $k_i = 2$ and two with $k_i = 1$. A trio of half-sibs sharing one parent can be produced in two different ways, both of which result in $\Sigma(k_i) = 6$ and $\sum\left(k_i^2\right) = 12$ (see text). In total, 40 different parents are required to produce the depicted relationships (one with $k_i = 3$, eight with $k_i = 2$, and 31 with $k_i = 1$, which yields $\Sigma(k_i) = 2S = 50$, $\sum\left(k_i^2\right) = 72$, and $N_eI = 111$ (from eqn 2)).

N_e calculated using PwoP (eqn 2). In total, 40 different parents are required to produce the relationships depicted in Fig. 1 (1 parent with $k_i = 3$, 8 with $k_i = 2$, and 31 with $k_i = 1$). These parental contributions produce $\Sigma(k_i) = 2S = 50$, $\sum\left(k_i^2\right) = 72$, and $N_e = 111$ (from eqn 2).

We have developed a program (coded in PYTHON 2.6.4, and available from the authors on request) that constructs the vector of k_i values and calculates N_e using eqn (2), based on an input file that specifies the relatedness category for each pair of progeny in the sample. The program creates a parental generation with $2S$ parents as empty groups. Target progeny are considered one at a time, and after evaluation, each is assigned to two distinct parents. Parents are excluded as possibilities if (i) their group contains non-siblings or (ii) another parental group would match more siblings. A half-sibling pair occurs together in only one parental group, while full siblings share both parental assignments. After all progeny are considered, the vector of k_i values is computed as the vector of parental group sizes.

Precision and bias

The analyses aforementioned are based on an optimal scenario: all progeny have been sampled, and all pairwise relationships among the progeny have been ascertained without error. Under these circumstances, a unique solution can be found for the terms $\Sigma(k_i)$ and $\sum\left(k_i^2\right)$, which allows one to calculate N_e using eqn (2). In most practical applications involving parentage analysis, effective size can only be estimated, in which case it is important to consider two major sources of uncertainty associated with PwoP:

1 Uncertainty associated with sampling only a fraction of the total number of progeny produced;
2 Uncertainty associated with imperfectly resolving the relationship categories.

Sampling error

To focus on the first issue (random errors associated with sampling progeny), we assume that relationship categories are identified correctly and that an unknown number of progeny are produced, but we have only sampled S of them. The true effective size would be revealed if we could sample all progeny. An important question thus becomes, ''What is the relationship between an estimate of N_e and true N_e, when the estimate is computed from a *sample* of progeny using PwoP?'' Intuitively, two *a priori* considerations suggest that such an estimate might be subject to bias: (i) the analysis leading to eqn (2) used the parametric formula for variance rather than the finite-sample formula and (ii) the vector of k_i values generated by PwoP based on only a sample of progeny will not only fail to include null parents, but it will also fail to include any parents that actually did produce offspring but whose offspring (by chance) did not appear in the sample.

To empirically evaluate these factors, we used a Wright–Fisher model (discrete generations, constant N, random mating but no selfing, each parent with an equal opportunity to produce offspring) to simulate production of S progeny whose parents were randomly chosen from N potential parents. As no selfing was allowed, we used a slight modification to eqn (1) that is appropriate for species that have separate sexes or are monoecious but avoid selfing:

$$N_e = \frac{\bar{k}N - 2}{\bar{k} - 1 + V_k/\bar{k}} \qquad \text{(eqn 5)}$$

(Crow & Denniston 1988, Equation 2). Making the substitutions above for $\bar{k}$ and V_k leads to (inbreeding N_e, without selfing)

$$N_e = \frac{2S - 2}{\frac{\sum\left(k_i^2\right)}{2S} - 1} \qquad \text{(eqn 6)}$$

Results of these simulations (Table 1), which can be considered 'optimal PwoP' because they assume the vector of k_i values can be assembled without error, provide important information about both bias and precision. First, analysing only a sample of progeny has virtually no systematic effect on $\hat{N}_e$ computed by PwoP: harmonic mean $\hat{N}_e$ was very close to the nominal N regardless of the number of progeny sampled. We also verified through simulations that essentially unbiased estimates are also obtained from PwoP using eqn (6), when sexes are separate and sex ratio is skewed (data not shown, but see Fig. 4). Second, unless N is large and S is small, PwoP estimates have relatively small coefficients of variation (CV). For example, a CV ≤ 0.35 can be achieved for N = 50–100 with $S \geq 25$ and for N = 500–1000 with $S \geq 100$, and much smaller CVs are possible with larger samples. However, if N is large and S is small, $\hat{N}_e$ from PwoP will have very wide confidence intervals (note very large CVs for $S = 25$, $N = 500$ and S = 25–50, $N = 1000$; Table 1). In these scenarios, the distribution of $\hat{N}_e$ is highly skewed toward large values. This occurs because, as fewer and fewer siblings are identified, most of the k_i values are 1. As a result, $\sum\left(k_i^2\right)$ approaches $\sum(k_i)$, and the denominator of eqn (2) approaches zero, leading to a large $\hat{N}_e$. In the limit, if all progeny are unrelated, this implies that each parent produces exactly one offspring, and $\hat{N}_e$ using eqns (2) or (6) becomes infinitely large. This makes biological sense, as finding no related individuals is the expected result when sampling progeny from an

Table 1 Assessment of bias and precision in estimates of N_e based on samples of progeny analysed with PwoP. This analysis assumes sibship reconstruction can be done without error ('Optimal PwoP'). Values shown are the harmonic mean $\hat{N}_e$ (and coefficient of variation of $\hat{N}_e$) for 10000 simulated Wright–Fisher populations, where parents of S progeny were randomly chosen from N potential parents. No selfing was allowed, so $\hat{N}_e$ was calculated using eqn (6)

Sample size (S)	Parental population (N)			
	50	100	500	1000
25	49.3 (0.21)	98.7 (0.35)	483.7 (550)	1026.2 (454)
50	49.4 (0.1)	100.0 (0.14)	494.4 (0.44)	1018.1 (84)
100	49.7 (0.05)	99.7 (0.07)	501.6 (0.17)	994.2 (0.26)
200	49.9 (0.02)	99.7 (0.03)	498.3 (0.08)	996.4 (0.11)

infinite number of parents. In the simulations, these infinite $\hat{N}_e$ values were recorded as 10^6 to simplify calculations of CV($\hat{N}_e$). Therefore, the exact CV values for these scenarios should be interpreted with caution; the salient point is that, if S is too small compared to N_e, the distribution of $\hat{N}_e$ can be highly skewed toward large (and potentially infinite) values. A similar result is found for other estimators of contemporary N_e (discussed by Waples & Do 2010).

It is useful to consider the probability of obtaining an estimate of $\hat{N}_e = \infty$ based on PwoP. If $\hat{N}_e = \infty$, the S progeny must have $2S$ unique parents. If we assume a Wright–Fisher model with random selfing, this outcome requires that $2S$ random draws, with replacement, have been made from N potential parents without having any parent drawn more than once. The probability of this occurring is $N/N \times (N-1)/N \times (N-2)/N \ldots \times (N-2S+1)/N$, which can be written as

$$\text{Prob(no sibs)} = \frac{N!}{(N-2S)!N^{2S}} \qquad \text{(eqn 7)}$$

If the population is randomly mating but not ideal (i.e., some individuals have a higher probability of being a parent than others), eqn (7) should still be valid if N_e is substituted for N.

Table 2 shows results of applying eqn 7 to various combinations of S and N_e. It is apparent that in many circumstances, only modest sample sizes are required before the probability of an infinite estimate becomes very low. For example, with true $N_e = 100$, an infinite $\hat{N}_e$ is very unlikely if 15 or more progeny have been sampled, and with $N_e = 200$, an infinite $\hat{N}_e$ is even more unlikely if at least 25 progeny have been sampled. However, with $N_e = 1000$, finding no relatives will be quite common in samples of 25 progeny or less, and the probability of an infinite $\hat{N}_e$ does not drop below 1% until sample size is almost 50 progeny. These results explain the high CV values shown in Table 1 for some simulation scenarios. With only 25 progeny sampled, we expect the fraction of $\hat{N}_e = \infty$ estimates to be 8% for $N_e = 500$ and 29% for $N_e = 1000$. The high CV for $S = 50$, $N_e = 1000$ was also influenced by ~0.6% infinite estimates (recorded as 10^6).

Table 2 Probability (from eqn 7) of finding no related individuals (and hence an infinite estimate of N_e) in a random sample of S progeny produced by $N_e = N$ effective parents, assuming random selfing. Last column shows the maximum finite $\hat{N}_e$ from eqn (8) for each sample size

N	S	Probability (no sibs)	Maximum finite $\hat{N}_e$
50	5	0.382	45
50	10	0.012	190
50	15	$<10^{-4}$	435
100	10	0.130	190
100	15	0.008	435
100	20	$<10^{-3}$	780
200	10	0.374	190
200	15	0.101	435
200	20	0.015	780
200	25	0.001	1225
500	15	0.412	435
500	25	0.079	1225
500	30	0.025	1770
500	40	0.001	3160
500	50	$<10^{-4}$	4950
1000	15	0.644	435
1000	25	0.288	1225
1000	40	0.039	3160
1000	50	0.006	4950
1000	60	$<10^{-3}$	7140

It is also easy to calculate the upper bound for finite $\hat{N}_e$, which occur if exactly one related pair of progeny is identified. With a single pair of half-sibs, the vector of k_i values includes $2S-2$ parents with $k_i = 1$ and one parent with $k_i = 2$. This leads to $\sum(k_i) = 2S$ and $\sum\left(k_i^2\right) = 2S + 2$, and eqn (2) becomes

$$\text{Maximum finite } \hat{N}_e = \frac{2S-1}{\frac{2S+2}{2S}-1} = 2S^2 - S \qquad \text{(eqn 8)}$$

Table 2 shows results of applying eqn (8) to the combinations of S and N values discussed earlier. One noteworthy result is that if sample size is very small (S = 5–15, depending on true N_e), the maximum finite $\hat{N}_e$ can be less than the true N_e. In that case, $\hat{N}_e$ is unreliable even though the harmonic mean is essentially unbiased, because the distribution has a mix of finite estimates less than the true N_e balanced by a fraction of infinite estimates. If true N_e is relatively large, then larger samples of progeny are needed to produce a more balanced distribution of potential $\hat{N}_e$ values. For example, for $S = 25$, the largest possible finite estimate of N_e is 1225 (Table 2),

which is more than twice as large as 500 but only marginally higher than 1000. If true N_e is 1000, therefore, sampling only 25 individuals would make it impossible for the distribution of $\hat{N}_e$ to be clustered around the true value.

Finally, we used simulated data to compare the distribution of $\hat{N}_e$ values from optimal PwoP with those obtained using a single-sample method for estimating N_e based on linkage disequilibrium (LD). To generate genetic data for the LD method, we used EasyPop (Balloux 2001) to simulate a Wright–Fisher model with equal numbers of each sex. We generated 'microsatellite' data for independent gene loci using a mutation rate of 5×10^{-4} and a *K*-allele model with a maximum of 10 allelic states and initiated the simulation using the Maximal diversity option. After running the simulation for enough generations (≥25) to produce average heterozygosities in the range (~0.8) typically found in many natural populations, we sampled genotypes from *S* progeny and used these to estimate N_e using the program LDNE (Waples & Do 2008). LDNE implements a bias correction (Waples 2006) to the standard Hill (1981) method; we report results after excluding alleles with frequency <0.02 (as suggested by Waples & Do 2010). PwoP estimates were generated as in Table 1 using the same *S* and *N* values, but using eqn (6) because the genetic data were generated with separate sexes. We considered two scenarios, with true N_e = 100 or 1000. The sample size (S = 50) and number of loci (20) used are comparable to those used in many contemporary studies of effective size in natural populations.

As shown in Fig. 2, the distribution of $\hat{N}_e$ from PwoP was much narrower than for the LD method. This was especially true for N_e = 1000, in which case PwoP reduced by a factor of nearly four the fraction of estimates that were more than twice the true N_e (from nearly 40% for the LD method to just over 10% for PwoP; Fig. 2A). Even with true N_e = 100, for which LD estimates are relatively robust, the distribution of PwoP estimates was less biased and substantially tighter (harmonic mean $\hat{N}_e$ = 99.7 (range 64–171) for PwoP compared to 113.9 (63–376) for LD; Fig. 2B).

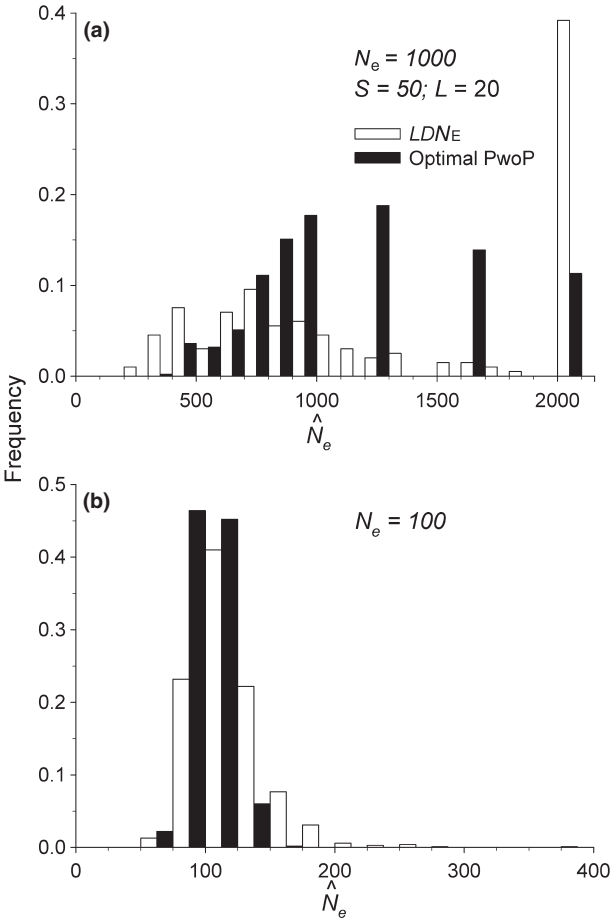

Fig. 2 Distribution of $\hat{N}_e$ for simulated Wright–Fisher populations with true N_e = 1000 (a) or 100 (b) using two methods. LDNE used L = 20 'microsat' loci, and optimal PwoP assumed the vector of k_i values was created without error. Both used samples of S = 50 individuals. In (a), the last bin on the right includes all estimates >2000.

Errors in sibship reconstruction

Evaluations of precision and bias in the previous section represent best-case ('optimal') scenarios for PwoP, because they assumed that sibships were accurately determined. Strictly speaking, this assumption is not required for the aforementioned results to be accurate, because the validity of eqn (2) depends only on the term $\sum(k_i^2)$ and not the exact pattern of relationship or the entire vector of k_i values. It would be possible, for example, for some errors to occur in sibship reconstruction that do not affect $\sum(k_i^2)$ and hence do not affect $\hat{N}_e$ from PwoP. Nevertheless, sibship reconstruction without parental genotypes as a guide is such a challenging task that uncertainty in this step will have a large influence on practical utility of PwoP.

A rigorous evaluation of this topic should consider two related issues: (i) limited power to resolve relationship categories given finite samples of progeny and gene loci and (ii) effects of genotypic errors. These two factors interact to affect bias and precision of sibship reconstructions, and they also involve tradeoffs that differ qualitatively depending on the markers used. For example, each microsatellite locus typically has many more alleles and hence more power for sibship reconstruction or parentage analysis than does a single single-nucleotide polymorphism (SNP) marker, but microsatellites also typically have higher genotyping error rates. Unless the

per-locus genotyping error rate is very low, as the number of loci increases the chances than any given individual will have at least one mis-scored genotype can become quite high, which can bias results even as theoretical power increases (e.g., see Anderson & Garza 2006).

We will not attempt a comprehensive assessment of this complex topic here. However, we want to provide some indication of the performance of PwoP that can be achieved with currently available software. Jones *et al.* (2010) listed seven freely available programs that attempt sibship reconstruction, but most of these consider only full sibs or partial subsets of half-sibs. Only two of the programs [COLONY (Jones & Wang 2010) and ML-RELATE (Kalinowski *et al.* 2006)] attempt the generalized sibship reconstruction envisioned by PwoP. Of these, COLONY in theory should provide more robust results, as it jointly considers the likelihood of larger patterns of relationship, whereas ML-RELATE independently determines the relationship of each pair of progeny. However, COLONY is computationally intensive and not easily adapted for simulation studies, while ML-RELATE is simple and quick and can read simulated genetic data in standard formats. Accordingly, we conducted a few exploratory runs in which ML-RELATE was challenged with simulated genotypic data (using EasyPop as described above), and the resulting determinations of pairwise relationships were used to calculate $\hat{N}_e$ using PwoP. ML-RELATE uses maximum likelihood to independently find the most likely relationship category for each progeny pair (parent-offspring, full sibling, half-sibling, unrelated). Because we simulated discrete-generation data, our samples of progeny had no parent-offspring dyads, so any parent-offspring determinations by ML-RELATE were scored according to the next most likely relationship category (which most frequently was half-sib).

Figure 3 shows results of analyses comparable to those in Fig. 2B (20 'microsat' loci, S = 50, N_e = 50, 100, 200). Under these conditions, precision using PwoP and ML-RELATE was actually quite high, but substantial biases were apparent. For example, with true N_e = 100, uncertainty in sibship reconstruction led to a sharp downward bias in ML-RELATE estimates (harmonic mean $\hat{N}_{e(\mathrm{ML-R})}$ = 78.5; all estimates fell in the range 65–98). The downward bias was more pronounced for true N_e = 200 (harmonic mean $\hat{N}_{e(\mathrm{ML-R})}$ = 101.8; range = 78–123), but for N_e = 50, the bias was modest and positive (harmonic mean $\hat{N}_{e(\mathrm{ML-R})}$ = 56.8; range = 48–70). These results suggest an interaction between sample size and effective size with respect to bias of $\hat{N}_e$ using ML-RELATE.

A likely explanation for this downward bias is overestimation of the number of pairs of progeny that are related, which would inflate the estimate of $\sum(k_i^2)$ and lead to an underestimate of N_e. Note that this can occur

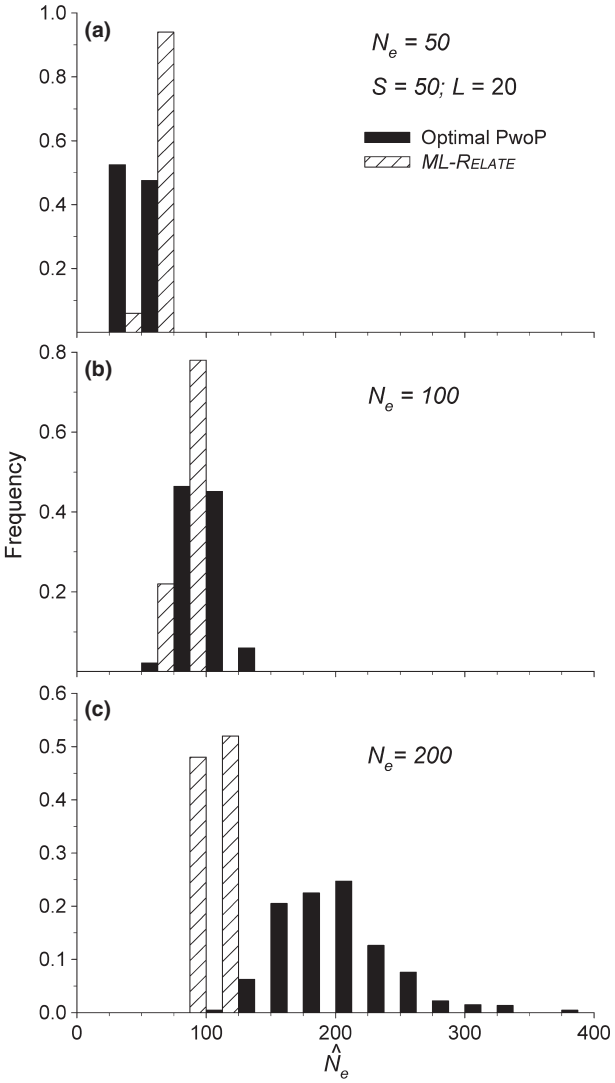

Fig. 3 Distribution of $\hat{N}_e$ for PwoP estimates when true N_e was 50 (a), 100 (b) or 200 ideal individuals (c). Each panel compares estimates based on (i) simulated demographic data under optimal conditions (assuming perfect sibship reconstruction, as in Fig. 2; black bars) and (ii) simulated genotypic data (using 20 'microsat loci', as in Fig. 2 for LDNE) analysed using the program ML-RELATE (hatched bars).

even if the probability of making a Type I error (wrongly inferring too high a relationship category) is lower than the probability of a Type II error (wrongly inferring too low a relationship category)—as reported previously for ML-RELATE (Kozfkay *et al.* 2008) and for another relationship estimation program (Thomas & Hill 2002). Unless N is very small, truly unrelated pairs will comprise most of the progeny, so even a low Type I error rate can lead to net estimates of more siblings than actually exist. We believe that this factor was responsible for the downward bias in $\hat{N}_e$ for N = 100 and 200. Although sibships were

also generally overestimated for $N = 50$, another factor appears to have been relatively more important in that case: impossible combinations of relationships that can occur as a result of relying on strictly pairwise analyses (e.g., A and B are determined to be full sibs, as are A and C, but B and C are not). The simulations with $N = 50$ had a relatively high frequency of these impossible combinations, which neither ML-RELATE nor our simple algorithm were designed to try to resolve. In our simple algorithm, these incompatibilities hinder the formation of large sibling groups under a single parent; this reduces the fraction of large k_i values [and reduces $\sum(k_i^2)$] and tends to inflate the estimate of N_e.

To evaluate the effects of marker type, we repeated the aforementioned analyses with true $N_e = 100$ and either 100 or 200 diallelic ('SNP') loci instead of 20 'microsat' loci. One hundred SNP loci produced results comparable to those for the 'microsat' data: precise but downwardly biased estimates using ML-RELATE (harmonic mean $\hat{N}_e = 74.8$; range 62–92). However, use of 200 SNPs led to precise estimates with relatively little downward bias (harmonic mean $\hat{N}_e = 93.4$; range 78–123) (Fig. 4A).

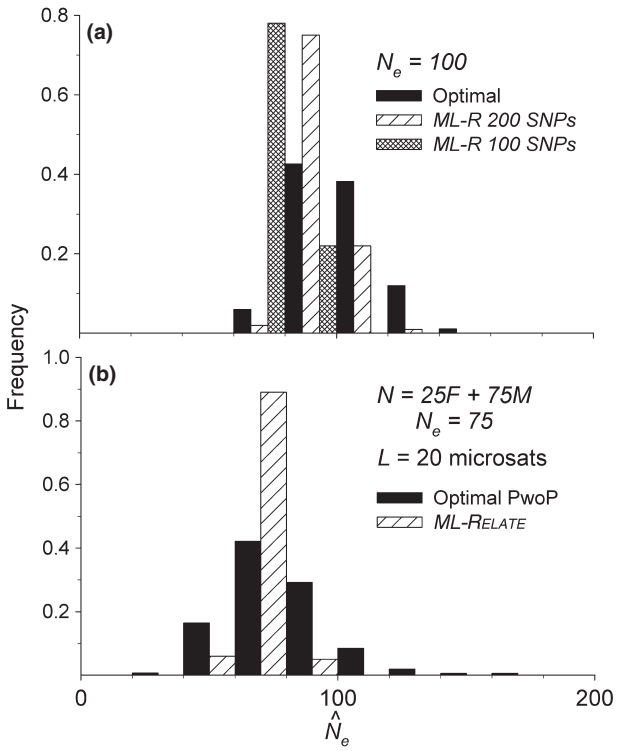

Fig. 4 (a) As in Figure 3, but using either 100 or 200 diallelic 'single-nucleotide polymorphism' loci for the ML-RELATE estimates. True N_e was 100. (b) Distribution of $\hat{N}_e$ for 'optimal' PwoP estimates when sex ratio was skewed (25 females + 75 males; true $N_e = 75$). ML-RELATE estimates used 20 'microsat' loci. In both panels, sample size was $S = 50$.

Simulation results presented so far have used ideal Wright–Fisher populations with equal sex ratio. Equations 5 and 6 account for skewed sex ratio, as overall V_k increases if the numbers of each sex are not equal, and Fig. 4B shows that 'optimal' PwoP estimates accurately reflect the lowered N_e from a 3:1 sex ratio (harmonic mean $\hat{N}_e = 75.7$; range 40–150). Skewed sex ratio also did not adversely affect estimates based on the program ML-RELATE: harmonic mean $\hat{N}_e$ was only slightly lower than expected (68.9), and the range of estimates (53–82) was actually tighter than under optimal PwoP (Fig. 4B).

Discussion

The new formulas (eqns 2 and 6) show that inbreeding effective size can be expressed in terms of a single unknown parameter ($\sum(k_i^2)$), independent of the total number of parents, N [of course, it is necessary to know or be able to estimate N if one is interested in the N_e/N ratio]. Except for the special case with $\bar{k} = 2$, this property is not shared by variance effective size (which depends on the number of progeny sampled), and this emphasizes the point that inbreeding N_e is the more useful measure of effective size for parentage analysis. The advantage of the new formulation is that it allows unbiased calculation of N_e in parentage analysis under a wide range of circumstances. Although the analyses considered here assumed that no parents can be genotyped, the method can also be applied when some but not all potential parents can be identified and sampled—a situation that occurs quite often in studies of natural populations (e.g., Emery *et al.* 2001; Araki *et al.* 2007).

Although we have not attempted a rigorous performance evaluation of PwoP, results of the analysis of simulated data establish two major points:

1 PwoP can provide unbiased and precise estimates of N_e from random samples of progeny, provided the vector of k_i values can be constructed accurately.
2 Accomplishing the latter will be challenging, and substantial biases can occur if systematic errors occur in reconstructing sibling relationships.

Fortunately, to estimate N_e using PwoP it is not necessary to reconstruct parental genotypes, which is exceedingly challenging unless at least some parents can be genotyped. The analyses described earlier depend only on reconstruction of sibling relationships, which can be estimated using currently available software. It should be possible to improve considerably on the biases indicated in Fig. 3 by adopting more sophisticated methods that jointly consider relationships among groups of related individuals. Performance with 200 'SNP' loci was encouraging (Fig. 4A); however, results obtained by Santure

et al. (2010) and others caution that use of large numbers of markers will not necessarily achieve the desired level of precision or accuracy. Use of fractional or probabilistic relationship assignments might also be an effective strategy for reducing bias and increasing precision, as could inclusion of other types of information, such as individual phenotypes (Walling *et al.* 2010).

Although the approach outlined here used sibship reconstructions to infer the vector of k_i values, that also is not necessarily required, as the key parameter to estimate is $\sum\left(k_i^2\right)$ and estimating the full vector of parental contributions is only an intermediate step in the process. This suggests that a more profitable approach might be to develop a method to specifically estimate $\sum\left(k_i^2\right)$, either directly or jointly with estimating the sibling relationships, as in full probability parentage analysis (Jones *et al.* 2010; Serbezov et al. 2010).

PwoP has some obvious parallels to Wang's (2009) sibship method for estimating N_e. Both use information from sibship reconstructions to estimate effective size; the difference is that Wang's method uses the sibship results to calculate the probability of different relationship categories under different hypotheses for N_e, while PwoP uses the sibship results to calculate the parental contributions and hence N_e directly using a demographic formula.

Future research might also focus on comparison of the performance of PwoP with other single-sample estimators besides LDNe (Pudovkin *et al.* 1996; Nomura 2008; Tallmon *et al.* 2008; Wang 2009) under a variety of realistic conditions. As discussed by Waples & Do (2010), calculating a combined estimate of effective size based on results of multiple methods can substantially increase precision (and potentially reduce bias), especially if the methods provide independent information about effective size. For example, Waples (1991) reported that $\hat{N}_e$ from the temporal and LD method are essentially uncorrelated, but comparable evaluations have not been performed for the different single-sample estimators.

Unlike some other genetic methods for estimating N_e, PwoP does not depend on the assumption of selective neutrality, as it relies on genetic data only to reconstruct parental contributions, nor does it depend on assumptions about the mating system. Like other estimators of contemporary N_e (and unlike estimators of long-term N_e), PwoP does not require an estimate of mutation rate, and mutation poses a problem only insofar as it might affect sibship reconstruction. Whether immigration (or other factors that might cause individuals from multiple populations to appear in the sample) represents a problem depends on the objectives and the quantity one is trying to estimate. Unlike some other estimators, PwoP does not depend on theoretical expectations for genetic processes in closed populations. If individuals from more than one population appear in the sample, the immigrants presumably would be determined to be unrelated to local individuals, which would tend to reduce $\sum\left(k_i^2\right)$ and increase $\hat{N}_e$. From one point of view, this would accurately reflect the reality that the sample is produced by more parents than occur in just the local population. On the other hand, this could be misleading if the goal is to estimate just the local N_e. In the latter case, it might be possible to use genetic assignment methods to exclude immigrants and focus only on locally produced progeny (as suggested by Wang 2009).

Although we did not evaluate this directly, PwoP presumably shares with most or all other estimators of contemporary N_e a sensitivity to sampling from age-structured populations. The standard formulas for effective size that PwoP is based on (eqns 1 and 5) assume discrete generations. If the sample is from a single cohort in an age-structured population, the estimate should be directly interpretable in terms of N_b, the effective number of breeders in one year or one breeding season. However, if mixed-age samples are taken from iteroparous species with overlapping generations, the resulting estimate can be difficult to interpret in terms of effective size for a generation as a whole (N_e; see Waples & Yokota 2007). More research is needed to better elucidate the relationship between N_b and N_e in age-structured species.

Acknowledgements

We are indebted to Mark Bravington for sharing unpublished data and to him, James Crow, Peter Smouse, and Ian Wilson for stimulating discussions. Three anonymous reviewers provided useful comments on an earlier draft. This work arose from a project funded by the Gordon and Betty Moore Foundation via the University of California, Santa Barbara and was inspired by code written by Eli Meir from SimBiotic Software.

Conflict of interest

The authors have no conflict of interest to declare and note that the sponsors of the issue had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Anderson EC, Garza JC (2006) The power of single-nucleotide polymorphisms for large-scale parentage inference. *Genetics*, **172**, 2567–2582.

Araki HA, Waples RS, Ardren WR, Cooper B, Blouin MS (2007) Effective population size of steelhead trout: influence of variance in reproductive success, hatchery programs, and genetic compensation between life-history forms. *Molecular Ecology*, **16**, 953–966.

Balloux F (2001) EasyPop (version 1.7): a computer program for population genetics simulations. *Journal of Heredity*, **92**, 301–302.

Caballero A (1994) Developments in the prediction of effective population size. *Heredity*, **73**, 657–679.

Crow JF (1954) Breeding structure of populations. II. Effective breeding number. In: *Statistics and Mathematics in Biology* (eds Kempthorne O,

Bancroft TA, Gowen JW, Lush JL), pp. 543–556. Iowa State College Press, Ames, IA.

Crow JF, Denniston C (1988) Inbreeding and variance effective population numbers. *Evolution*, **42**, 482–495.

Crow JF, Morton NE (1955) Measurement of gene frequency drift in small populations. *Evolution*, **9**, 202–214.

DeWoody JA, Walker D, Avise JC (2000) Genetic parentage in large half-sib clutches: theoretical estimates and empirical appraisals. *Genetics*, **154**, 1907–1912.

Emery AM, Wilson IJ, Craig S, Boyle PR, Noble LR (2001) Assignment of paternity groups without access to parental genotypes: multiple mating and developmental plasticity in squid. *Molecular Ecology*, **10**, 1265–1278.

Eriksson A, Mehlig B, Panova M, Andre C, Johannesson K (2010) Multiple paternity: determining the minimum number of sires of a large brood. *Molecular Ecology Resources*, **10**, 282–291.

Hill WG (1981) Estimation of effective population size from data on linkage disequilibrium. *Genetical Research*, **38**, 209–216.

Jones AG (2005) GERUD2.0: a computer program for the reconstruction of parental genotypes from progeny arrays with known or unknown parents. *Molecular Ecology Notes*, **5**, 708–711.

Jones AG, Avise JC (1997) Polygynandry in the dusky pipefish *Syngnathus floridae* revealed by microsatellite DNA markers. *Evolution*, **51**, 1611–1622.

Jones AG, Small CM, Paczolt KA, Ratterma NL (2010) A practical guide to methods of parentage analysis. *Molecular Ecology Resources*, **10**, 6–30.

Jones OR, Wang J (2010) COLONY: a program for parentage and sibship inference from multilocus genotypic data. *Molecular Ecology Resources*, **10**, 551–555.

Kalinowski ST, Wagner AP, Taper ML (2006) ML-RELATE: a computer program for maximum likelihood estimation of relatedness and relationship. *Molecular Ecology Notes*, **6**, 576–579.

Kozfkay CC, Campbell MR, Heindel JA *et al*. (2008) A genetic evaluation of relatedness for broodstock management of captive, endangered Snake River sockeye salmon, *Oncorhynchus nerka*. *Conservation Genetics*, **9**, 1421–1430.

Myers EM, Zamudio KR (2004) Multiple paternity in an aggregate breeding amphibian: the effect of reproductive skew on estimates of male reproductive success. *Molecular Ecology*, **13**, 1951–1963.

Nomura T (2008) Estimation of effective number of breeders from molecular coancestry of single cohort sample. *Evolutionary Applications*, **1**, 462–474.

Pudovkin AI, Zaykin DV, Hedgecock D (1996) On the potential for estimating the effective number of breeders from heterozygote-excess in progeny. *Genetics*, **144**, 383–387.

Santure AW, Stapley J, Ball AD, Birkhead TR, Burke T, Slate J (2010) On the use of large marker panels to estimate inbreeding and relatedness: empirical and simulation studies of a pedigreed zebra finch population typed at 771 SNPs. *Molecular Ecology*, **19**, 1439–1451.

Serbezov D, Bernatchez L, Olsen EM, Vøllestad LA (2010) Mating patterns and determinants of individual reproductive success in brown trout (*Salmo trutta*) revealed by parentage analysis of an entire stream living population. *Molecular Ecology*, **19**, 3193–3205.

Tallmon DA, Koyuk A, Luikart G, Beaumont MA (2008) ONESAMP: a program to estimate effective population size using approximate Bayesian computation. *Molecular Ecology Resources*, **8**, 299–301.

Tatarenkov A, Healy CIM, Grether GF, Avise JC (2008) Pronounced reproductive skew in a natural population of green swordtails, *Xiphophorus helleri*. *Molecular Ecology*, **20**, 4522–4534.

Thomas SC, Hill WG (2002) Sibship reconstruction in hierarchical population structures using Markov chain Monte Carlo techniques. *Genetical Research*, **79**, 227–234.

Walling CA, Pemberton JM, Hadfield JD, Kruuk LEB (2010) Comparing parentage inference software: reanalysis of a red deer pedigree. *Molecular Ecology*, **19**, 1914–1928.

Wang JL (2004) Sibship reconstruction from genetic data with typing errors. *Genetics*, **166**, 1963–1979.

Wang J (2009) A new method for estimating effective population size from a single sample of multilocus genotypes. *Molecular Ecology*, **18**, 2148–2164.

Waples RS (1991) Genetic methods for estimating the effective size of cetacean populations. *Report of the International Whaling Commission*, (special issue 13), 279–300.

Waples RS (2002) Evaluating the effect of stage-specific survivorship on the N_e/N ratio. *Molecular Ecology*, **11**, 1029–1037.

Waples RS (2006) A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. *Conservation Genetics*, **7**, 167–184.

Waples RS, Do C (2008) LDNE: a program for estimating effective population size from data on linkage disequilibrium. *Molecular Ecology Resources*, **8**, 753–756.

Waples RS, Do C (2010) Linkage disequilibrium estimates of contemporary N_e using highly variable genetic markers: a largely untapped resource for applied conservation and evolution. *Evolutionary Applications*, **3**, 244–262.

Waples RS, Yokota M (2007) Temporal estimates of effective population size in species with overlapping generations. *Genetics*, **175**, 219–233.