Journal of Heredity 2009:100(1):106–113
doi:10.1093/jhered/esn088
Advance Access publication October 20, 2008

Simultaneous Estimation of Null Alleles and Inbreeding Coefficients

Igor J. Chybicki and Jaroslaw Burczyk

From the Department of Genetics, Institute of Experimental Biology, Kazimierz Wielki University, Bydgoszcz 85-064, Poland.

Address correspondence to I. J. Chybicki at the address above, or e-mail: igorchy@ukw.edu.pl.

Abstract

Although microsatellites are a very efficient tool for many population genetics applications, they may occasionally produce "null" alleles, which, when present in high proportion, may affect estimates of key parameters such as inbreeding and relatedness coefficients or measures of genetic differentiation. In order to account for the presence of null alleles, it is first necessary to estimate their frequency within studied populations. However, the commonly used null allele frequency estimators are not of general applicability because they can produce upwardly biased estimates when a population under study experiences some inbreeding. In such a case, 2 formerly described approaches, population inbreeding model and individual inbreeding model, can be applied for simultaneous estimation of null allele frequencies and of the inbreeding coefficient. In this study, we demonstrate the properties and utility of these 2 methods and show that they outperform the commonly used approaches in the estimation of null allele frequencies based on genotypic data. The methods are applied to empirical data from a natural population of European beech (*Fagus sylvatica* L.), and results are briefly discussed. The methods presented in this paper are implemented in the Windows-based user-friendly INEST computer program (available free of charge at http://genetyka.ukw.edu.pl/INEst10_setup.exe).

Key words: *estimation, Fagus sylvatica, inbreeding, microsatellite, null allele, SSR*

Microsatellites are widely used genetic markers suitable for many population genetic applications. They allow to assess genetic differentiation among populations (Goldstein et al. 1995; Slatkin 1995), as well as to track gene flow by means of parentage analysis (e.g., Asuka et al. 2005; Bacles et al. 2005). Because microsatellite markers are primarily neutral (although they can be linked to loci subjected to selection, e.g., Kohn et al. 2000), they are often used for estimating mating system parameters, such as inbreeding or relatedness within population (Blouin 2003). Although microsatellites have many advantages for population genetics studies, including high polymorphism, ease of molecular analysis, and interpretation, genotyping of microsatellite markers requires careful examination because of relatively frequent occurrence of null alleles (reviewed in Dakin and Avis 2004). Null alleles have been reported for microsatellites in numerous organisms, for example, humans (Callen et al. 1993), oystercatcher (Van Treuren 1998), weevil (Liewlaksaneeyanawin et al. 2002), including tree species such as pine (Vogl et al. 2002), spruce (Nascimento de Sousa et al. 2005), and common ash (Bacles et al. 2005). They usually originate from a point mutation in flanking regions of a microsatellite, which cannot be observed without sequencing of these regions (Holm et al. 2001; Vornam et al. 2004).

The recessive character of null alleles makes the assessment of individual genotypes uncertain in the case of the heterozygous state with a normally amplifying allele. Hence, in the presence of substantial proportion of null alleles at a particular locus, the observed heterozygosity would be largely underestimated. Consequently, null alleles affect especially population parameter estimates, which are based on the proportion of heterozygotes. Especially, when calculating Wright's inbreeding coefficient F based on microsatellite data, it remains unclear to what extent the estimated F reflects the actual level of inbreeding in the studied population and to what degree it was affected by the presence of null. Null alleles can also affect genetic differentiation measures, causing the more overestimation in F_{ST} the lower actual gene flow is (Chapuis and Estoup 2006). Additionally, the presence of null alleles, if not accounted for, leads to large error rates in parentage assignment and subsequent substantial bias in inferred mating system, behavior, and dispersal parameters (Dakin and Avis 2004).

Null alleles may be detected experimentally in several ways. For example, the presence of null alleles might be inferred from results of controlled crossing (Pastorelli et al. 2003). However, this method is inefficient if one intends to assess the frequency of null alleles in a population because null alleles are, in general, randomly distributed across genotypes. At the population level, the excess of homozygotes at a particular locus might be a sign of possible null allele occurrence, especially when it is not concordant with the biology of the species. In such a situation, redesigning

primers could help in answering whether the increased homozygosity is a result of null alleles or other factors (e.g., inbreeding, Wahlund effect) (Holm et al. 2001; Vornam et al. 2004). This approach, however, requires much effort and might be insufficient in particular situations because the newly designed primer regions could still be weakly conserved in the genome. Alternatively, within-population null allele frequency can be estimated using information available from genotypic proportions (Chakraborty et al. 1992; Brookfield 1996; Chapuis and Estoup 2006; Kalinowski and Taper 2006). As long as a population is at Hardy–Weinberg equilibrium, these methods provide good estimates of null allele frequency (Chapuis and Estoup 2006; Kalinowski and Taper 2006). However, when a population experiences inbreeding, leading to an excess of homozygotes, they may substantially overestimate null allele frequencies (Van Oosterhout et al. 2006).

The attempts have been already undertaken to estimate null allele frequencies within nonrandomly mating populations, when the deviation from the Hardy–Weinberg equilibrium is due to mating between relatives. To our knowledge, so far 2 different approaches have been proposed to estimate null allele frequency in such a case of inbreeding within populations. Both approaches are basically founded on the same population genetic model, here called the inbred population model (see Materials and methods); however, neither has been used in practice. The main difference between the 2 methods concerns the treatment of the inbreeding coefficient in describing the inbreeding within a population. The first approach, developed originally for the ABO blood group allelic system (Yasuda 1968), describes the inbreeding by one summary parameter, that is, the average within-population inbreeding coefficient. On the contrary, the second approach employs the individual inbreeding measures, which in detail depicts the inbreeding of each individual within the population (Vogl et al. 2002). Both methods can be used to estimate null allele frequencies with simultaneous estimation of inbreeding as a separate parameters, using relevant statistical approaches (Schull and Ito 1969; Vogl et al. 2002; this study). However, little is known about the properties of these estimators. For example, the efficiency of the population inbreeding approach was studied only for the ABO data. It was shown that ABO phenotypic proportions do not give sufficient information to jointly estimate allele frequencies and the inbreeding coefficient (Yasuda 1968). Later, Schull and Ito (1969) concluded that, at least theoretically, simultaneous estimation of the recessive allele frequency and the inbreeding coefficient is possible unless the actual inbreeding is close to zero. However, how the population inbreeding approach behaves when several loci are used at the same time has not yet been evaluated. Contrary to the population inbreeding model (PIM)–based method, nothing is known about the statistical properties of the individual inbreeding approach, which was used for empirical data sets only (Vogl et al. 2002; Muir et al. 2004).

It is worth noting that a PIM-based estimator of the null allele frequency in a nonequilibrium population was recently proposed (Van Oosterhout et al. 2006). Unfortunately, this method requires independent information on the inbreeding coefficient and therefore is not fully applicable to cases when such estimates are not available.

In the present paper, we adapt the Yasuda's (1968) model to the case of a multilocus and multiallelic analysis. This allows us to use data from several loci at the same time in order to estimate simultaneously null allele frequencies at each locus and the average level of the intrapopulation inbreeding as a multilocus parameter. Using extensive computer simulations, we explore the statistical properties of both the PIM-based estimators and the individual inbreeding model (IIM)–based estimators. Additionally, the methods are applied to an empirical example data set obtained for a natural population of *Fagus sylvatica* L., and results and their implications are briefly discussed. Appropriate Windows-based computer software was developed to make the method available (available free of charge at http://genetyka.ukw.edu.pl/INEst10_setup.exe).

Materials and Methods

For clarity, it is important to note that the inbreeding coefficient F referred to in the next section has to be considered strictly as a probability that the 2 alleles at a locus are identical by descent (IBD) (Malécot 1948), which implicates that such F parameter must fall into a 0–1 interval. Notably, the above-mentioned definition differs from the one originally made by Wright (1922) (i.e., F as a correlation of alleles within individuals).

Inbred Population Model

Let us consider a sample of individuals randomly drawn from a population where each individual is genotyped at a number of loci. At every locus, there are 2 classes of alleles: dominant (i.e., visible and distinguishable; within this class, m alleles are mutually codominant, i.e., $A_1, A_2, \ldots A_m$) and recessive ones (for simplification, we will consider only one recessive allele per locus, A_0, a null). Therefore, genotypes are unobservable in certain cases, and instead, only individual phenotypes are known precisely. At a given locus, each individual phenotype falls into 1 of 3 following categories: 1) dominant for the j-th ($j = 1, 2, \ldots, m$) allele (i.e., with only one allele distinguishable, A_{j-}), 2) a heterozygote for the j-th ($j = 1, 2, \ldots, m$) and the k-th ($k\ [\neq j] = 1, 2, \ldots, m$) dominant allele ($A_j, A_k$), or 3) a homozygote for a recessive allele (A_0, A_0). Within the category (1), an individual can have A_{j-} phenotype because either it is homozygous for the j-th dominant allele or it is heterozygous for the j-th dominant allele and a recessive allele. Given p_{ij} being the frequency of the j-th allele at the i-th locus, a homozygote for the j-th allele at the i-th locus can be observed either by chance, with the probability $p_{ij}^2(1 - F)$, or due to inbreeding, with probability $p_{ij}F$. Given p_{i0} being the frequency of a recessive allele at the i-th locus, the heterozygote for the j-th dominant allele and a recessive allele

at i-th locus is expected with the probability $2p_{ij}p_{i0}(1-F)$. Thus, the total probability that an individual has the phenotype A_{j-} is equal to $p_{ij}^2 + 2p_{ij}p_{i0}(1-F) + p_{ij}(1-p_{ij})F$. An individual can have A_jA_k $(j \neq k)$ phenotype, with the probability $2p_{ij}p_{ik}(1-F)$. Finally, at i-th locus, an individual can have the A_0A_0 phenotype (recessive homozygote) either by chance, with probability $p_{i0}^2(1-F)$, or due to inbreeding, with the probability $p_{i0}F$. Thus, the total probability that an individual has the phenotype A_0A_0 is equal to $p_{i0}^2 + p_{i0}(1-p_{i0})F$.

The model presented above can be considered in 2 variants: 1) when F means the population-wide average inbreeding coefficient (PIM), that is, the probability that 2 alleles at a random locus are IBD in a randomly chosen individual in a population or 2) when F means the individual inbreeding coefficient (F_i) (IIM), corresponding to the probability that the 2 alleles at a random locus in the i-th individual are IBD.

Regardless of which variant is considered, the model presented above matches the scenario when a number of individuals drawn from a population are genotyped at a number of microsatellite loci, having besides normally amplifying alleles (corresponding to dominant alleles in the model) also nonamplifying "null" ones (corresponding to recessive alleles in the model). Then, given phenotypic observations, one can use the model in order to develop maximum likelihood (ML) estimators of allele frequencies (p_{ij}, including a null allele—p_{i0}) and of the inbreeding coefficient F. Depending on the model variant, allele frequencies and the inbreeding coefficient can be estimated following the 2 alternative approaches described below.

PIM-based Estimator

Let n_{ij-}, n_{ijk}, and n_{i00} represent sample counts of individuals having at the i-th locus phenotype A_{j-}, A_jA_k, and A_0A_0, respectively. When individuals are typed at l unlinked loci, then, assuming drawing with replacement, the probability of phenotypes has the multinomial distribution (see Weir 1996), leading to the likelihood function:

$$L(n|F,p) \propto c \times \prod_{i}^{l} \left\{ \prod_{j}^{m_i} \left[p_{ij}^2 + 2p_{ij}p_{i0}(1-F) + p_{ij}(1-p_{ij})F \right] \right\}^{n_{ij-}} \times \times \prod_{j,k;j\neq k}^{m_i} \left[2p_{ij}p_{ik}(1-F) \right]^{n_{ijk}} \times \left[p_{i0}^2 + p_{i0}(1-p_{i0})F \right]^{n_{i00}} \quad (1)$$

where n—phenotypic counts, p—allele frequencies, m_i—number of dominant (nonnull, i.e., visible) alleles at the i-th locus, and c—constant dependent on the phenotypic counts in a sample. Although closed-form ML solutions for p and F cannot be formulated, given observed phenotypic counts, the likelihood 1) can be maximized numerically providing the ML estimates of allele frequencies and the inbreeding coefficient.

It is worth noting that setting $F = 0$ in Equation 1 reduces it to the well-known ML estimator of allele frequencies (including nulls), introduced and studied independently by Chapuis and Estoup (2006) and Kalinowski and Taper (2006). To differentiate this special case method from the one described in this study, the former will be referred to hereafter as the random mating model (RMM)–based estimator and the latter as the PIM one.

IIM-based Estimator

Although this method basically relies on the inbred population model, a closed form of the likelihood formula for the IIM is difficult to obtain. Instead, a set of conditional distributions can be formulated, thus capturing the relationship between the individual inbreeding coefficients (F) and the population allele frequencies (p) (Vogl et al. 2002). Then, using the Gibbs sampler technique, one can obtain the approximate full posterior distribution, which can be used to compute estimates of allele frequencies and individual inbreeding coefficients, given observed data. The detailed description of the method is beyond the scope of this paper. However, readers interested in the formulation of this method should refer to the original 2002 paper of Vogl et al. (2002). Here, we only note that Equation 5 in Vogl et al. (2002) is incorrect and should be replaced by $\frac{F_i p_{lm} + (1-F_i)p_{lm}^2}{F_i p_{lm} + (1-F_i)p_{lm}^2 + (1-F_i)2p_{lm}p_{l0}}$ (using the original notation, see Vogl et al. 2002).

Simulation Study

In order to evaluate the accuracy and precision of the methods described above, computer simulations were carried out. Our simulation algorithm was chosen to mimic, although in a very simplified way, the reproduction process existing within a large population of an annual plant mating either at random (with probability t) or via self-fertilization (with probability $1-t$). Such a population, after a number of generations, exhibits the average inbreeding, which stabilizes at the level $F = (1-t)/(1+t)$. The detailed simulation algorithm is as follows (Coelho and Vencovsky 2003, with modifications): 1) specify outcrossing rate (t) given the expected inbreeding coefficient according to the equation $t = (1-F)/(1+F)$; 2) randomly generate k allele frequencies at L loci; 3) attribute to each of the L loci of the N individuals a given genotype as a function of allele frequencies; 4) draw an individual (i); 5) generate a random number from a (0,1) uniform distribution (x); 6) if $x < t$, draw a second individual (j), otherwise take $j = i$; 7) for each locus, take one allele at random from the i-th and the j-th individual's genotype and combine them to form the genotype of a progeny; 8) go to step 4 N times to obtain N individuals representing the next generation (to neglect the effect of random genetic drift here $N = 10\,000$); 9) repeat steps 3–8 for the desired number of generations (here 20); and 10) draw a sample of S individuals from the last generation and convert their genotypes so that all heterozygotes for the null and the l-th alleles appear phenotypically as homozygotes for the l-th allele. The sample of S individuals generated in this manner was used to estimate parameters of interest using both the PIM and the IIM methods. In the PIM, allele frequencies and the inbreeding coefficient were estimated using Expectation-Maximization algorithm based on Equation 1 (see Supplementary Material). Using the IIM, the parameters were estimated using the Gibbs sampler (Vogl et al. 2002), after

Table 1. Bias and RMSE of estimators of null allele frequencies and inbreeding coefficient based on simulated data

				Null alleles			Visible alleles			Inbreeding coefficient	
S	*L*	*F*		RMM	PIM	IIM	RMM	PIM	IIM	PIM	IIM
50	5	0	Bias	−0.004	−0.023	−0.006	0.000	0.002	0.001	0.047	0.005
			RMSE	0.023	0.043	0.024	0.006	0.008	0.006	0.081	0.013
		0.1	Bias	0.035	−0.007	0.002	−0.004	0.001	0.000	0.009	−0.020
			RMSE	0.045	0.033	0.027	0.008	0.008	0.007	0.084	0.067
		0.2	Bias	0.073	−0.001	−0.003	−0.008	0.000	0.000	−0.008	−0.003
			RMSE	0.078	0.037	0.031	0.012	0.009	0.009	0.094	0.071
	10	0	Bias	0.000	−0.010	−0.001	0.000	0.001	0.000	0.026	0.002
			RMSE	0.022	0.027	0.022	0.006	0.006	0.006	0.046	0.003
		0.1	Bias	0.037	−0.001	−0.003	−0.004	0.000	0.000	−0.006	0.000
			RMSE	0.045	0.030	0.025	0.009	0.007	0.008	0.064	0.054
		0.2	Bias	0.073	−0.004	−0.008	−0.008	0.001	0.001	−0.009	0.003
			RMSE	0.079	0.030	0.025	0.012	0.009	0.009	0.072	0.054
100	5	0	Bias	−0.004	−0.022	−0.007	0.000	0.002	0.001	0.042	0.005
			RMSE	0.015	0.033	0.016	0.004	0.005	0.004	0.063	0.007
		0.1	Bias	0.040	−0.003	0.001	−0.004	0.000	0.000	0.002	−0.008
			RMSE	0.045	0.028	0.022	0.007	0.005	0.005	0.061	0.045
		0.2	Bias	0.079	−0.001	−0.005	−0.008	0.000	0.001	−0.005	0.006
			RMSE	0.083	0.032	0.023	0.011	0.007	0.006	0.069	0.046
	10	0	Bias	−0.001	−0.014	−0.003	0.000	0.002	0.000	0.032	0.002
			RMSE	0.014	0.024	0.014	0.004	0.004	0.004	0.047	0.004
		0.1	Bias	0.041	−0.002	−0.002	−0.004	0.000	0.000	0.005	0.005
			RMSE	0.046	0.024	0.018	0.007	0.006	0.006	0.049	0.038
		0.2	Bias	0.078	−0.003	−0.009	−0.008	0.000	0.001	0.000	0.015
			RMSE	0.082	0.021	0.018	0.011	0.006	0.006	0.045	0.033

S, sample size; *F*, actual average inbreeding coefficient; *L*, number of loci. Results for 3 estimators: the RMM, the PIM based, and the IIM based. Bias and RMSE for null allele frequencies averaged over *L* loci. Bias and RMSE for visible allele frequencies averaged over *L* loci and alleles.

10 000 iterations, in addition with 1000 iterations as a burn-in step. The simulation algorithm as well as both estimation methods were implemented in an OBJECT PASCAL/DELPHI computer program.

In the simulation study, we were interested particularly in the 3 factors, which presumably influence the accuracy and precision of the estimator, which are: 1) sample size ($S = 50$, 100), 2) number of loci ($L = 5$, 10), and 3) actual inbreeding coefficient ($F = 0$, 0.1, 0.2) (Table 1). Each locus displays 10 alleles. Frequencies of alleles were generated randomly so that they varied in each repetition of the simulation. However, special attention was paid to null allele frequencies because they are not expected to have a uniform distribution (Dakin and Avis 2004). In this study, the prior distribution of null allele frequencies was close to the empirical distribution published in the review paper of Dakin and Avis (2004) (Figure 1). Each simulation scenario was repeated 1000 times, and finally bias and root mean squared error (RMSE) of estimates were computed according to the following equations:

$$\text{Bias} = \text{Avg}(\hat{\theta}_i - \theta_i), \tag{2}$$

$$\text{RMSE} = \sqrt{\text{Avg}\left[(\hat{\theta}_i - \theta_i)^2\right]}, \tag{3}$$

where $\hat{\theta}_i$ is an estimate of a given parameter (a null allele frequency at a given locus, a visible [normally amplifying] allele frequency, or inbreeding coefficient) and θ_i an actual value of a given parameter in the *i*-th repetition of the simulation, $\text{Avg}\left(\sum_i \hat{\theta}_i - \theta_i\right)$ indicates arithmetic mean over all repetitions.

In order to evaluate the methods based on the inbred population model, we compared them with the method that assumes random mating. For this purpose, we chose the ML estimator (Chapuis and Estoup 2006; Kalinowski and Taper 2006, here called RMM) because it appeared to be the most accurate and precise among all estimators currently available (Chapuis and Estoup 2006; Kalinowski and Taper 2006). The RMM method was applied to the data generated based on the same simulation scheme.

Application to Empirical Data: A Natural Population of European Beech

As an empirical example, 104 adult individuals from an European beech population (Chybicki J, Trojankiewicz M, Oleska A, Dzialuk A, Burczyk J, unpublished data) genotyped for 9 SSR loci (Table 2; for details, see Supplementary Materials) were used. The population under investigation was sampled in the Northwestern part of Poland, near Bobolice, 40 km Southeast from Koszalin. In this region, European beech forms a natural monospecific rich forest, characteristic of the morainic landscape. The stand selected for this study is a natural reserve, where beech is the main focus of protection. The sampled stand covers about 9.8 ha and is surrounded by other European beech forests and European beech–Norway spruce mixed stands. Little is known about the history of this stand. However, the

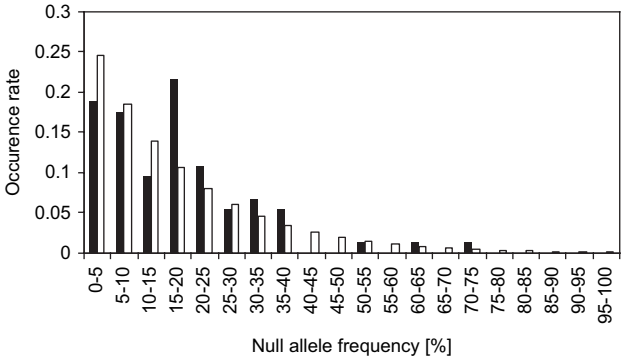

Figure 1. Empirical (black boxes) and best-fitting exponential (open boxes) distributions of null allele frequencies based on published data (Dakin and Avis 2004). The exponential distribution was used as a prior distribution in the simulation study, providing the same average null allele frequency as the empirical distribution (0.175).

presence of very old beech trees scattered across the stand and its surroundings suggests that the beech forest persisted in this area for several generations. Therefore, the population may be considered as natural and stable in terms of the ecosystem dynamics.

European beech (*F. sylvatica* L.) is a late-successional forest tree species, growing generally in a temperate climate. It covers a large continuous geographic range in Europe. European beech is a wind-pollinated tree, mating mostly through outcrossing (Merzeau et al. 1994; Rossi et al. 1996; Wang 2003), although controlled pollination experiments showed that beech is capable for self-fertilization at a rate up to 13% of total matings (Nielsen and Schaffalitzky De Muckadell 1954). Many populations of European beech showed an evident excess of homozygotes (Cuguen et al. 1988; Comps et al. 1990, 1991; Beletti and Lanteri 1996; Leonardi and Menozzi 1996; Hazler et al. 1997; Vornam et al. 2004; Jump and Peñuelas 2007), which was attributed particularly to the effect of inbreeding or, more generally, to the isolation by distance (Cuguen et al. 1988; Gömöry et al. 1999). Because beechnuts have no structures facilitating dispersal, they often germinate beneath a mother tree. Therefore, in naturally regenerating beech populations, there is an obvious opportunity for the development of a primary half-sib family structure (Gregorius et al. 1986; see also Asuka et al. 2005). Furthermore, assuming that pollination occurs to some degree between neighboring trees (Wang 2004), this may result in consanguineous matings in consecutive generations. In consequence, there is a high opportunity for a development of the genetic structure typical for the isolation by distance phenomena, including elevated F coefficient and strong clustering of relatives (Vornam et al. 2004; Jump and Peñuelas 2007).

Null alleles at high frequencies have been detected for some microsatellites used in this study (Pastorelli et al. 2003), and they can be present also in the studied population given that some loci showed significant excess of homozygotes (Table 2). However, because beech is known for increased inbreeding in natural populations, application of any genetic models working under the a priori assumption of random mating does not seem relevant for this species. With either of the 2 methods presented in this paper (i.e., the PIM based and the IIM based), we can relax the assumption of random mating in order to estimate null allele frequencies, accounting for the (unknown) level of inbreeding within the population.

Table 2. Parameters of genetic structure of the studied population

Locus	k	H_e	H_o	P	$null_{PIM}$	$null_{IIM}$
fs 1-15	9	0.756	0.721	0.135	0.019	0.022
fs 1-25	14	0.760	0.485	0.000	0.221^{**}	0.220^{**}
fs 3-04	3	0.336	0.365	0.872	0.000	0.010
fs 4-46	17	0.838	0.550	0.000	0.202^{**}	0.201^{**}
mfc 5	15	0.816	0.702	0.008	0.072^{*}	0.072^{*}
sfc 0007-2	5	0.699	0.712	0.655	0.000	0.009
sfc 0036	7	0.547	0.644	1.000	0.000	0.001
sfc 0161	14	0.733	0.712	0.299	0.004	0.003
sfc 1143	11	0.878	0.952	0.990	0.000	0.001
Mean	10.6	0.707	0.649	—	—	—

Null allele frequency estimates significantly different from zero indicated by asterisks: $^{*}P < 0.01$ and $^{**}P < 0.001$. k, number of alleles, H_e, expected heterozygosity, H_o, observed heterozygosity, P, P value of the exact test for heterozygote deficiency, $null_{PIM}$, estimate of null allele frequency given the PIM-based estimator, $null_{IIM}$, estimate of null allele frequency given the IIM-based estimator. The average inbreeding coefficient estimated simultaneously with null alleles using either the PIM or the IIM methods was equal to 0.

Results and Discussion

Simulation Study

The 3 following methods were evaluated by means of simulation: 1) RMM-based estimator (Chapuis and Estoup 2006; Kalinowski and Taper 2006), 2) PIM-based estimator (Yasuda 1968; this study), and 3) IIM-based estimator (Vogl et al. 2002). Simulations confirmed that the RMM method is not of general applicability because its efficiency strongly depends on the actual inbreeding coefficient. RMM is unbiased only when $F = 0$. In this case, the RMM method is very robust and, as shown in earlier studies, outperformed other methods (Chapuis and Estoup 2006; Kalinowski and Taper 2006). However, when $F > 0$, RMM provides upwardly biased null allele frequencies (Table 1, see also Van Oosterhout et al. 2006), with the bias reaching, on average, up to +0.08, when the actual $F = 0.2$. Given that the mean frequency of null alleles in simulated data was 0.175, RMM overestimated null allele frequencies about 1.5 times, on average. When comparing RMM to either the PIM or the IIM method, both PIM and IIM were found to be more, although not equally, independent on the actual inbreeding coefficient.

When the total spectrum of the actual F (0–0.2) is considered in simulations, IIM appeared to be the least biased estimator of both allele frequencies and the average inbreeding coefficient. PIM provided unbiased estimates as long as the actual F was strictly positive. On the contrary, when actual $F = 0$, PIM gave upwardly biased inbreeding coefficient estimates (up to +0.047) and downwardly biased null allele frequency estimates, with the largest bias equal to −0.023. In the case of the actual $F = 0$, the PIM method provided, however, less biased null allele frequencies, when compared with the RMM method for the actual $F > 0$. Therefore, application of the PIM estimator made some improvement in the assessment of null allele frequencies, although its quality depends to some degree on the actual inbreeding coefficient.

The RMSE is a function of both the bias and the variance of the estimator. Therefore, when inferring the precision of the estimators, one should carefully inspect bias first. Taking this into account, RMSE values indicated that the sample size influenced only precision of the estimators having no impact on their accuracy. Again, IIM was characterized by the lowest values of RMSE, regardless of the sample size, the number of loci, or the actual inbreeding coefficient. On average, IIM estimates of null allele frequencies had RMSE values that were about 2.2 times lower than RMM estimates and 1.4 times lower PIM estimates. Furthermore, only the IIM method provided relatively stable RMSE values for null allele estimates within the total spectrum of the actual F variable considered.

The number of loci influenced only the PIM and IIM methods because the RMM method is in fact a single-locus estimator. The number of loci affected both the accuracy and the precision of estimates with similar effects on both the PIM and the IIM methods. In general, the performance of the methods increased (low bias and low RMSE) when increasing the number of loci (Table 1). This phenomenon can be related to the improved ability of making a proper balance between the inbreeding and the presence of null alleles when more loci are used.

Simulations also showed that, regardless of the sample size, the number of loci, or the method used, the accuracy and precision of the estimates of visible allele frequencies remain consistently stable. In the case of PIM and IIM methods, they were also not much influenced by the actual F. On the contrary, the RMM estimates of visible allele frequencies were sensitive to the actual F. However, it is an indirect effect because the actual inbreeding coefficient influences primarily null allele frequency estimates, which as a result are overestimated, and consequently visible alleles are underestimated.

European Beech Example Study

Using the methods based on the inbred population model, we provided estimates of null allele frequencies within the studied population. The estimations showed that a large proportion of null alleles is present particularly at the 3 loci: *fs 1-25*, *fs 4-46*, and *mfc 5*, for which null allele frequency estimates converged to 0.22, 0.20, and 0.07, respectively (Table 2). In all 3 cases, they were significantly greater than zero. Interestingly, the 2 loci, *fs 1-25* and *fs 4-46*, indicated putative null homozygotes, which were observed through the single-locus—as well as the multiplex—polymerase chain reaction (PCR) analysis using 8 loci. Although including those individuals in the analysis was reasonable, we performed independently additional analyses excluding putative null homozygotes from the sample. In this case, null allele frequencies decreased slightly reaching 0.16 and 0.16 at *fs 1-25* and *fs 4-46*, respectively, however, not affecting other estimates, including the inbreeding coefficient. The observed reduction in null frequency estimates seems reasonable, as the excluded putative null homozygotes carried 10 and 8 copies of null allele at *fs 1-25* and *fs 4-46* loci, respectively. Among the remaining loci, null alleles were detected also for *fs 1-15*, but in frequencies not significantly different from zero. Regardless of the method used (i.e., PIM or IIM), such estimates of the population inbreeding coefficient were practically equal to zero and not significantly different one from another.

Our detailed analyses showed that the increased homozygosity observed in the studied beech population was due to the presence of null alleles and not to inbreeding, although the later one could be suspected given susceptibility to isolation by distance of this species (Cuguen et al. 1988; Gömöry et al. 1999; Vornam et al. 2004; Wang 2004). The analyzed microsatellite loci showed moderate polymorphism, with the average number of alleles per locus equal to 10.6 (Table 2). However, as indicated in the simulations, having 104 individuals genotyped at 9 loci with 10 alleles on average should provide very accurate results either by PIM or by IIM method. Therefore, the obtained estimates of both null allele frequencies and inbreeding coefficient are of high confidence.

Excess of homozygosity in a natural population of European beech detected in this study is in agreement with earlier studies of this species based on a similar set of microsatellite markers (Vornam et al. 2004; Jump and Peñuelas 2006; Buiteveld et al. 2007). However, unlike the earlier reports, this study showed that the excess of homozygotes was exclusively due to null alleles and not (even partly) to inbreeding. Although some SSR markers used in this study were transferred from the related species (i.e., *Fagus crenata*), it is worth noting that null alleles were detected particularly at SSR loci, which were developed for the species under study (*F. sylvatica*). Overall, when comparing 4 native markers (i.e., developed for *F. sylvatica*) with 5 nonnative ones (i.e., developed for *F. crenata*), the native markers showed on average a marked deficiency of heterozygotes ($H_o = 0.53$, $H_e = 0.67$), whereas the nonnative markers showed a slight excess of heterozygotes ($H_o = 0.74$, $H_e = 0.73$). It is possible that markers transferred from another species can, for some reasons, exhibit irregularities in PCR reaction, such as an amplification of nonspecific fragments. We cannot exclude that some of nonnative markers might be scored incorrectly, causing an excess in heterozygosity due to misinterpretation of PCR

products. It regarded especially the 2 loci: *sfc* 0036 and *sfc* 1143, which showed a high excess of heterozygotes (Table 2). In order to assess whether nonnative markers influence the estimates of null allele frequencies (indirectly, by deflating the inbreeding coefficient), we performed additional analysis based on the genotypes at native markers only. The estimated frequencies of null alleles did not change in a case of PIM method. On the other hand, null allele frequency estimates based on IIM dropped down from 0.220 to 0.206 in case of *fs 1-25* and from 0.201 to 0.185 in case of *fs 4-46*. This decrease in null allele frequencies was compensated with the estimate of the average inbreeding coefficient equaled 0.036. Nonetheless, it has to be stressed that F estimated with IIM method was not significantly different from zero (standard error = 0.033). Therefore, one might conclude that nonnative character of some SSR markers did not influence much the results.

General Remarks

The 2 methods described in this study work within the frames of the inbred population model; therefore, any violations of the model's assumptions can influence their accuracy. It should be noted that estimates of PIM and IIM methods might be slightly inaccurate due to the inherent properties of the estimators. As the inbreeding coefficients are estimated as a probability obeying a constraint (0,1) (Vogl et al. 2002), the F estimator may have a right-skewed distribution when the actual $F \ll 0.5$ or a left-skewed distribution when the actual $F \gg 0.5$. The larger variance attributed to the actual inbreeding coefficient (e.g., resulted from the interindividual variation in outcrossing rate, see Coelho and Vencovsky 2003) the more skewed estimator for marginal values of the actual F should be expected. The PIM method seems particularly subjected to this property, which to some degree was anticipated by Schull and Ito (1969). However, simulations showed that increasing a number of loci to 10 or more would reduce the bias of the PIM method substantially.

It should be noticed that in our study, we assumed that a lack of PCR products for a given locus is due to homozygous state of null allele (see Equation 1). Although in our empirical example this reasoning was justified, as mentioned earlier, in real world data, it is often unsure whether a source of lacking data is null allele presence or some artifact during PCR reaction or scoring procedure. However, as pointed by Kalinowski and Taper (2006), the random lack of data could be parameterized and easily incorporated into the likelihood function (Vogl et al. 2002; Kalinowski and Taper 2006). As such, it would counterbalance in particular null allele effect, and when not accounted for the random lack of data, it would inflate null allele frequency estimates (Kalinowski and Taper 2006).

In summary, although both the PIM and the IIM methods outperform the RMM method, when there is some uncertainty about the actual inbreeding within a population, IIM seems the best choice for empirical studies. Among all the methods, it demonstrated the highest accuracy and precision. Moreover, the IIM method has the additional advantage of making possible inferences on the distribution of individual inbreeding coefficients and, indirectly, about the mating system within the population (Vogl et al. 2002; Muir et al. 2004). Nonetheless, additional studies are needed to assess the utility of IIM based in this field.

Availability

The methods presented in this paper are implemented in the Windows-based user-friendly INEST computer program (available free of charge at http://genetyka.ukw.edu.pl/INEst10_setup.exe).

Supplementary Material

Supplementary materials can be found at http://www.jhered.oxfordjournals.org/.

Funding

Polish Ministry of Science and Higher Education (2 P06L 039 30) to J.B.

Acknowledgments

We are grateful to Cecile Bacles for helpful discussions, comments, and suggestions that improved the manuscript. We thank our colleagues from the Department of Genetics UKW for their help with fieldwork.

References

Asuka Y, Tomaru N, Munehara Y, Tani N, Tsumura Y, Yamamoto S. 2005. Half-sib family structure of *Fagus crenata* saplings in an old-growth beech–dwarf bamboo forest. Mol Ecol. 14:2565–2575.

Bacles CFE, Burczyk J, Lowe AJ, Ennos RA. 2005. Historical and contemporary mating patterns in remnant population of the forest tree *Fraxinus excelsior* L. Evolution. 59:979–990.

Beletti P, Lanteri S. 1996. Allozyme variation among European beech (*Fagus sylvatica* L) stands in Piedmont, North-Western Italy. Silvae Genet. 45:33–37.

Blouin MS. 2003. DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. Trends Ecol Evol. 18:503–511.

Brookfield JFY. 1996. A simple new method for estimating null allele frequency from heterozygote deficiency. Mol Ecol. 5:453–455.

Buiteveld J, Vendramin GG, Leonardi S, Kamer K, Geburek T. 2007. Genetic diversity and differentiation in European Beech (*Fagus sylvatica* L) stands varying in management history. For Ecol Manage. 247:98–106.

Callen DF, Thompson AD, Shen Y, Phillips HA, Richards RI, Mulley JC, Sutherland GR. 1993. Incidence and origin of 'null' alleles in the (AC)n microsatellite markers. Am J Hum Genet. 52:922–927.

Chakraborty R, De Andrade M, Daiger SP, Budowle B. 1992. Apparent heterozygote deficiencies observed in DNA typing and their implications in forensic applications. Ann Hum Genet. 56:45–57.

Chapuis M-P, Estoup A. 2006. Microsatellite null alleles and estimation of population differentiation. Mol Biol Evol. 24:621–631.

Coelho ASG, Vencovsky R. 2003. Intrapopulation fixation index dynamics in finite populations with variable outcrossing rates. Sci Agric. 60:305–313.

Comps B, Thiébaut B, Paule L, Merzeau D, Letouzey J. 1990. Allozymic variability in beechwoods (*Fagus sylvatica* L) over Europe: spatial differentiation among and within populations. Heredity 65:407–417.

Comps B, Thiébaut B, Sugar I, Trinajstic I, Plazibat M. 1991. Genetic variation of the Croatian beech stands (*Fagus sylvatica* L): spatial differentiation in connection with the environment. Ann Sci For. 48:15–28.

Cuguen J, Merzeau D, Thiebaut B. 1988. Genetic structure of the European beech stands (*Fagus sylvatica* L): F-statistics and importance of mating system characteristics in their evolution. Heredity 60:91–100.

Dakin EE, Avis JC. 2004. Microsatellite null alleles in parentage analysis. Heredity 93:504–509.

Goldstein DB, Linares AR, Cavalli-Sforza LL, Feldman MW. 1995. An evaluation of genetic distances for use with microsatellite loci. Genetics. 139:463–471.

Gömöry D, Paule L, Brus R, Zhalev P, Tomović Z, Gračan J. 1999. Genetic differentiation and phylogeny of beech on the Balkan Peninsula. J Evol Biol. 12:746–754.

Gregorius H-R, Krauhausen J, Muller-Starck G. 1986. Spatial and temporal genetic differentiation among the seed in a stand of *Fagus sylvatica* L. Heredity 57:255–262.

Hazler K, Comps B, Šugar I, Melovski L, Tashev A, Gračan J. 1997. Genetic structure of *Fagus sylvatica* L populations in southeastern Europe. Silvae Genet. 46:229–236.

Holm L-E, Loeschcke V, Bendixen C. 2001. Elucidation of the molecular basis of a null allele in a Rainbow trout microsatellite. Mar Biotechnol. 3:555–560.

Jump AS, Peñuelas J. 2006. Genetic effects of chronic habitat fragmentation in a wind-pollinated tree. Proc Natl Acad Sci USA. 103:8096–8100.

Jump AS, Peñuelas J. 2007. Extensive spatial genetic structure revealed by AFLP but not SSR molecular markers in the wind-pollinated tree, *Fagus sylvatica*. Mol Ecol. 16:925–936.

Kalinowski ST, Taper ML. 2006. Maximum likelihood estimation of the frequency of null alleles at microsatellite loci. Conserv Genet. 7:991–995.

Kohn MH, Pelz H-J, Wayne RK. 2000. Natural selection mapping of the warfarin-resistance gene. Proc Natl Acad Sci USA. 97:7911–7915.

Leonardi S, Menozzi P. 1996. Spatial structure of genetic variability in natural stands of *Fagus sylvatica* L (beech) in Italy. Heredity 77:359–368.

Liewlaksaneeyanawin C, Ritland CE, El-Kassaby YA. 2002. Inheritance of null alleles for microsatellites in the White pine weevil (*Pissodes strobi* [Peck] [Coleoptera: Curculionidae]). J Hered. 93:67–70.

Malécot G. 1948. Les mathématiques de l'hérédité. Paris (France): Masson & Cie.

Merzeau D, Comps B, Thiébaut B, Letouzey J. 1994. Estimation of *Fagus sylvatica* L mating system parameters in natural populations. Ann Sci For. 51:163–173.

Muir G, Lowe AJ, Fleming CC, Vogl C. 2004. High nuclear genetic diversity, high levels of outcrossing and low differentiation among remnant populations of *Quercus petraea* at the margin of its range in Ireland. Ann Bot. 93:691–697.

Nascimento de Sousa S, Finkeldey R, Gailing O. 2005. Experimental verification of microsatellite null alleles in Norway spruce (*Picea abies* [L] Karst.): implications for population genetic studies. Plant Mol Biol Rep. 23:113–119.

Nielsen PC, Schaffalitzky De Muckadell M. 1954. Flower observations and controlled pollinations in *Fagus*. Silvae Genet. 3:6–17.

Pastorelli R, Smulders MJM, Van't Westende WPC, Vorman B, Giannini R, Vettori C, Vendramin GG. 2003. Characterization of microsatellite markers in *Fagus sylvatica* L and *Fagus orientalis* Lipsky. Mol Ecol Notes 3:76–78.

Rossi P, Vendramin GG, Giannini R. 1996. Estimation of mating system parameters in two Italian natural populations of *Fagus sylvatica*. Can J For Res. 26:1187–1192.

Schull WJ, Ito PK. 1969. A note on the estimation of the ABO gene frequencies and the coefficient of inbreeding. Am J Hum Genet. 21:168–170.

Slatkin M. 1995. A measure of population subdivision based on microsatellite allele frequencies. Genetics 139:457–462.

Van Oosterhout C, Weetman D, Hutchinson WF. 2006. Estimation and adjustment of microsatellite null alleles in nonequilibrium populations. Mol Ecol Notes 6:255–256.

Van Treuren R. 1998. Estimating null allele frequencies at a microsatellite locus in the oystercatcher (*Haematopus ostralegus*). Mol Ecol. 7:1413–1417.

Vogl C, Karhu A, Moran G, Savolainen O. 2002. High resolution analysis of mating systems: inbreeding in natural populations of *Pinus radiata*. J Evol Biol. 15:433–439.

Vornam B, Decarli N, Gailing O. 2004. Spatial distribution of genetic variation in a natural beech stand (*Fagus sylvatica* L) based on microsatellite markers. Conserv Genet. 5:561–570.

Wang KS. 2003. Relationship between empty seed and genetic factors in European beech (*Fagus sylvatica* L). Silva Fenn. 37:419–428.

Wang KS. 2004. Gene flow in European beech (*Fagus sylvatica* L). Genetica. 122:105–113.

Weir BS. 1996. Genetic data analysis II. Sunderland (MA): Sinauer.

Wright S. 1922. Coefficients of inbreeding and relationship. Am Nat. 56:330–338.

Yasuda N. 1968. Estimation of the inbreeding coefficient from phenotype frequencies by a method of maximum likelihood scoring. Biometrics 24:915–935.

Received October 3, 2007; Revised July 25, 2008; Accepted September 10, 2008

Corresponding Editor: Halina Knap