6-2012

Maximum Likelihood Estimation of Individual Inbreeding Coefficients and Null Allele Frequencies

Nathan Hall
Western Washington University

Laina Mercer
Western Washington University

Daisy Phillips
Western Washington University

Jonathan Shaw
North Carolina Wildlife Resources Commission

Amy D. Anderson
Western Washington University, amy.anderson@wwu.edu

Genet. Res., Camb. (2012), **94**, *pp.* 151–161.
doi:10.1017/S0016672312000341

Maximum likelihood estimation of individual inbreeding coefficients and null allele frequencies

NATHAN HALL[1]*, LAINA MERCER[1]†, DAISY PHILLIPS[1]‡, JONATHAN SHAW[2] AND AMY D. ANDERSON[1]§
[1] *Department of Mathematics, Western Washington University, Bellingham, WA 98225, USA*
[2] *North Carolina Wildlife Resources Commission, Raleigh, NC 27695, USA*

(*Received 6 January 2012; revised 12 May 2012; accepted 16 May 2012; first published online 18 July 2012*)

Summary

In this paper, we developed and compared several expectation–maximization (EM) algorithms to find maximum likelihood estimates of individual inbreeding coefficients using molecular marker information. The first method estimates the inbreeding coefficient for a single individual and assumes that allele frequencies are known without error. The second method jointly estimates inbreeding coefficients and allele frequencies for a set of individuals that have been genotyped at several loci. The third method generalizes the second method to include the case in which null alleles may be present. In particular, it is able to jointly estimate individual inbreeding coefficients and allele frequencies, including the frequencies of null alleles, and accounts for missing data. We compared our methods with several other estimation procedures using simulated data and found that our methods perform well. The maximum likelihood estimators consistently gave among the lowest root-mean-square-error (RMSE) of all the estimators that were compared. Our estimator that accounts for null alleles performed particularly well and was able to tease apart the effects of null alleles, randomly missing genotypes and differing degrees of inbreeding among members of the datasets we analysed. To illustrate the performance of our estimators, we analysed previously published datasets on mice (*Mus musculus*) and white-tailed deer (*Odocoileus virginianus*).

1. Introduction

An individual is said to be inbred if its parents are genetically related to each other. The degree to which an individual is inbred can be summarized by that individual's inbreeding coefficient, f, which can be defined as the probability that the individual's two alleles at a random autosomal locus will both be copies of a single allele present in one of the individual's ancestors. In other words, f is the probability that an individual's two alleles will be identical by descent (IBD) from some ancestor.

There has been recent interest in the estimation of inbreeding coefficients or the related topic of estimating relatedness between individuals in the presence of inbreeding (Anderson & Weir, 2007; Purcell *et al.*, 2007; Chybicki & Burczyk, 2009; Wang, 2011*a*, *b*; Yang *et al.*, 2011). The main topic of this paper is the development of maximum likelihood estimators for estimating inbreeding coefficients from molecular marker data.

The presence of null alleles complicates the estimation of inbreeding coefficients. A null allele is an allele that cannot be directly observed during genotyping. An individual homozygous for a null allele will not produce any readable alleles at that marker and so will be recorded as having a missing genotype. An individual heterozygous for the null allele and another allele will be recorded as being homozygous for that other allele. The uncertainty in genotypes caused by the presence of null alleles can cause difficulty in interpreting the results of a number of kinds of population genetic studies (Chapuis & Estoup, 2007; Chybicki & Burczyk, 2009) and there has been recent interest in detecting null alleles and/or estimating their frequencies (Kalinowski & Taper, 2006; Van Oosterhout *et al.*, 2006; Chybicki & Burczyk,

* Current Address: Department of Mathematics, Whatcom Community College, Bellingham, WA 98226, USA
† Current Address: Seattle Children's Research Institute, Core for Biomedical Statistics, Seattle, WA 98105, USA
‡ Current Address: Department of Statistics, Pennsylvania State University, University Park, PA 16802, USA
§ Corresponding author: Department of Mathematics, Western Washington University, Bellingham, WA 98225, USA. E-mail: amy.anderson@wwu.edu

2009; Girard, 2011). In developing a method that could estimate inbreeding coefficients in the presence of null alleles, we noted that both null alleles and inbreeding cause an increase in observed homozygosity. If used effectively, though, a dataset consisting of genotypes of multiple individuals genotyped at several markers can be used to distinguish between null alleles and inbreeding because the two sources of excess homozygosity leave somewhat different footprints in the data. For example, markers with common null alleles will tend to have more missing data than other markers, especially among individuals who are more inbred. In contrast, at loci with no null alleles (or rare null alleles), inbred and non-inbred individuals should show similar rates of missingness. In addition, more highly inbred individuals will tend to have more missing genotypes (because such individuals are more likely to be homozygous for the null allele), particularly at markers with higher null allele frequencies. By looking at inbreeding and null alleles together, one can take advantage of these types of patterns in the data.

In this paper, we will compare new and previously existing methods that either (1) estimate individual inbreeding coefficients assuming known allele frequencies and no null alleles, (2) jointly estimate inbreeding coefficients and allele frequencies, again assuming no null alleles, or (3) jointly estimate individual inbreeding coefficients and allele frequencies in the presence of null alleles.

2. Theory and methods

(i) *Existing methods for estimating inbreeding coefficients when there are no null alleles*

Perhaps the most sophisticated method in this category is Leutenegger's method (Leutenegger *et al.*, 2003), which performs maximum likelihood estimation of inbreeding coefficients using (potentially) linked genetic markers. Leutenegger's method is the only method to be mentioned in this paper that models linkage between loci. Since we are concerned with unlinked markers, we did not include Leutenegger's method directly in our calculations. However, Leutenegger's method applied to unlinked markers (with no genotyping error) should be equivalent to our Method 1 described below. The only difference between the two is the algorithm used to numerically maximize the likelihood function.

Another estimator, which we call the 'Simple' estimator, is

$$\hat{f}_{\mathrm{Sim}} = 1 - \frac{H_O}{H_E}, \tag{1}$$

where H_O is the total number of loci at which the individual is heterozygous and H_E is the expected number of loci at which the individual is heterozygous, given the known allele frequencies. If p_{jk} is the relative frequency of allele k at marker j, m is the number of genotyped markers and n_j is the number of possible alleles at marker j, then

$$H_E = \sum_{j=1}^{m} \left(1 - \sum_{k=1}^{n_j} p_{jk}^2\right). \tag{2}$$

Note that this estimator is slightly different than Wright's population inbreeding coefficient estimator because Wright's estimator deals with genotypes from multiple individuals at a single marker, whereas this estimator deals with genotypes from a single individual at multiple markers. This estimator is similar in spirit to the one presented in Carothers *et al.* (2006), but differs in the details of the calculations. In the case in which there are only two possible alleles at a marker and allele frequencies are known (not estimated from the dataset), this estimator is the same as the estimator calculated by the PLINK program (Purcell *et al.*, 2007).

Another popular moment estimator, due to Ritland (specifically, eqn (5) in Ritland, 1996), is

$$\hat{f}_{\mathrm{Rit}} = \frac{\sum_{j=1}^{m} \sum_{k=1}^{n_j} \frac{S_{jk} - p_{jk}^2}{p_{jk}}}{\sum_{j=1}^{m} (n_j - 1)}, \tag{3}$$

where S_{jk} is 1 if the individual is homozygous for allele k at marker j and 0 otherwise. This estimator can be viewed as a variant of eqn (5) in Li & Horvitz (1953), except that Li's version assumes multiple individuals have been genotyped at a single locus, while Ritland's version applies to a single individual genotyped at multiple loci and, additionally, Li's version assumes allele frequencies have been estimated from the dataset by allele counting.

Both the Simple and Ritland estimators have the property that they do not constrain their estimated inbreeding coefficients to be between 0 and 1. If one considers only inbreeding coefficients between 0 and 1, one might replace any negative estimate by 0 and any estimate greater than 1 by 1 itself. The versions of the Simple and Ritland estimators for which we made these adjustments are called the adjusted Simple and adjusted Ritland estimators.

The above estimators do not take uncertainty in allele frequencies into account. Vogl *et al.* (2002) developed a Bayesian method to jointly estimate allele frequencies and inbreeding coefficients using a dataset containing multiple individuals genotyped at multiple markers. The Vogl model assumes that the inbreeding coefficients in the dataset are random values from a beta distribution. The final estimated inbreeding coefficients take both the beta prior distribution and the observed data into account. We ran Vogl's algorithm with the prior recommendation by Vogl *et al.*

(a beta prior with $\alpha_1=\alpha_2=0{\cdot}001$) and also using a uniform prior (a beta prior with $\alpha_1=\alpha_2=1{\cdot}0$).

(ii) *Existing method for estimating null allele frequencies, assuming no inbreeding*

Kalinowski & Taper (2006) produced an expectation–maximization (EM) algorithm to estimate allele frequencies and showed that it produced smaller root mean square errors (RMSE) than other estimators. The Kalinowski and Taper method considers one marker at a time and estimates the frequencies of all alleles at the marker (including a null allele) as well as a parameter, β, that measures the rate at which genotypes are randomly missing for reasons unrelated to null alleles. The β parameter is essential – without it, a marker with many missing genotypes but no other sign of excess homozygosity would produce a high estimated null allele frequency since, without β, the estimator would assume that any missing genotype is in fact homozygous for the null allele.

(iii) *Existing method for jointly estimating null allele frequencies and inbreeding coefficients*

Vogl *et al*. (2002), in their paper on estimating inbreeding coefficients, also described how their Bayesian method could be adapted to the case of null alleles.

(iv) *New Method 1: an EM algorithm for maximum likelihood estimation of inbreeding coefficients when allele frequencies are known*

When allele frequencies are known, individuals in the dataset can be considered one at a time. Since we are considering unlinked markers, the likelihood (the probability of the data, viewed as a function of the parameter f) is the product of the single-marker likelihoods. Let g_j be the genotype of the individual at marker j, let $\vec{g}=(g_1, \ldots, g_m)$ and let $A_{j1}, \ldots, A_{jn_j}$ be the possible alleles at marker j. Then the likelihood function is

$$L(f)=\prod_{j=1}^{m} \Pr(g_j|f,\vec{p}), \tag{4}$$

where, letting $\vec{p}$ denote the matrix of allele frequencies at the various markers, with p_{jk} denoting the relative frequency of allele k at marker j,

$$\Pr(g_j|f,\vec{p})=\begin{cases} fp_{jk}+(1-f)p_{jk}^2, & \text{if } g_j=A_{jk}A_{jk}, \\ 2(1-f)p_{jk}p_{jl}, & \text{if } g_j=A_{jk}A_{jl} \ (k\neq l). \end{cases} \tag{5}$$

The maximum likelihood estimate for f is found by maximizing the likelihood, $L(f)$, with respect to f.

We will perform this maximization via an EM algorithm. For each marker, j, define a random variable X_j as follows:

$$X_j=\begin{cases} 1, & \text{if the 2 alleles at marker } j \text{ are IBD} \\ 0, & \text{if the 2 alleles at marker } j \text{ are not IBD.} \end{cases} \tag{6}$$

If the X_j values were known, then the maximum likelihood estimator (MLE) for f would be $\hat{f}=\sum_{j=1}^{m} X_j/m$. If the X_j values were unknown but we could calculate their expected values (noting that $E[X_j|f,\vec{g}]$ is the probability that an individual's 2 alleles at marker j are IBD, given the individual's genotypes and inbreeding coefficient), the MLE for f would still be easy to find ($\hat{f}=\sum_{j=1}^{m} E[X_j|f,\vec{g}]/m$). Unfortunately, the probabilities $E[X_j|f,\vec{g}]$ depend on f and hence cannot be calculated without first knowing f. That leaves us with the following: if we knew f, we could estimate the $E[X_j|f,\vec{g}]$ values and, if we knew the $E[X_j|f,\vec{g}]$ values, we could estimate f. The EM algorithm proceeds by alternating between steps in which we estimate the $E[X_j|f,\vec{g}]$ values based on our current estimate for f, and steps in which we update our estimate for f based on our current values for $E[X_j|f,\vec{g}]$. The algorithm begins with an initial guess for f, call it $f^{(0)}$. Let $f^{(z)}$ denote our estimate for f at the start of the zth iteration of our algorithm. The steps for the $(z+1)$st iterations are:

1. Given $f^{(z)}$, estimate the value for X_j for every marker j using the following equation:

$$E_z\left[X_j|f^{(z)},\vec{g}\right]=\begin{cases} \frac{f^{(z)}p_{jk}}{f^{(z)}p_{jk}+(1-f^{(z)})p_{jk}^2}, & \text{if } g_j=A_{jk}A_{jk}, \\ 0, & \text{if } g_j=A_{jk}A_{jl} \ (k\neq l). \end{cases} \tag{7}$$

2. Given the $E_z\left[X_j|f^{(z)},\vec{g}\right]$ estimates, update the estimate for f using the following equation:

$$f^{(z+1)}=\frac{\sum_j E_z\left[X_j|f^{(z)},\vec{g}\right]}{m}. \tag{8}$$

Maximization is accomplished by iterating eqns (7) and (8) until convergence is reached. Note that the algorithm should be repeated using several different starting values for f because an EM algorithm can converge to a local rather than a global maximum (see, for example, Wu, 1983). A more formal derivation of this algorithm is included in the online Supplementary Material at http://journals.cambridge.org/GRH.

(v) *New Method 2: an EM algorithm for joint maximum likelihood estimation of inbreeding coefficients and allele frequencies*

We consider a dataset with N individuals genotyped at m markers. Let f_i denote the inbreeding coefficient

of individual i ($i=1, \ldots, N$) and g_{ij} be the genotype of individual i at marker j. Letting $\vec{f}=(f_1, \ldots, f_N)$ and $\vec{p}$ be the matrix of allele frequencies, the likelihood equation now becomes

$$L(\vec{f},\vec{p})=\prod_{i=1}^{N}\prod_{j=1}^{m}\Pr(g_{ij}|\vec{f},\vec{p}), \tag{9}$$

where the $\Pr(g_{ij}|\vec{f},\vec{p})$ values are still calculated as in eqn (5). The maximization of the likelihood is now taken with respect to $\vec{f}=(f_1, \ldots, f_N)$ and all the allele frequencies ($\vec{p}$) jointly. Let $\delta_{ijkl}=1$ if individual i's genotype at marker j is $A_{jk}A_{jl}$ and $\delta_{ijkl}=0$ otherwise and let X_{ij} be defined such that $X_{ij}=1$ if individual i's two alleles at marker j are IBD and $X_{ij}=0$ otherwise. Using $E_z[X_{ij}]$ as a shorthand notation for $E[X_{ij}|f_i^{(z)}, g_{ij}]$, the equations for the EM algorithm are

$$E_z[X_{ij}]=\begin{cases}\frac{f_i^{(z)}p_{jk}^{(z)}}{f_i^{(z)}p_{jk}^{(z)}+\left(1-f_i^{(z)}\right)\left(p_{jk}^{(z)}\right)^2}, & \text{if } g_{ij}=A_{jk}A_{jk},\\ 0, & \text{if } g_{ij}=A_{jk}A_{jl}\ (k\neq l),\end{cases} \tag{10}$$

$$p_{jk}^{(z+1)}=\frac{\sum_{i=1}^{N}\left[\delta_{ijkk}\left(E_z[X_{ij}]+2\left(1-E_z[X_{ij}]\right)\right)+\sum_{l\neq k}\delta_{ijkl}\right]}{\sum_{l=1}^{n_j}\sum_{i=1}^{N}\left[\delta_{ijll}\left(E_z[X_{ij}]+2\left(1-E_z[X_{ij}]\right)\right)+\sum_{h\neq l}\delta_{ijhl}\right]}, \tag{11}$$

$$f_i^{(z+1)}=\frac{\sum_j E_z[X_{ij}]}{m}. \tag{12}$$

Here, eqns (10) and (12) are analogous to eqns (7) and (8) in the previous method. Equation (11) provides updated estimates of the allele frequencies, given the estimated X_{ij} values. In spirit, the numerator in eqn (11) represents a calculation in which we go through the list of all individuals genotyped at marker j and, for each individual homozygous for allele A_k, we count 1 allele if the individual's 2 alleles are IBD and 2 distinct alleles otherwise. For each individual heterozygous for allele A_k, we count one allele. Since we cannot see whether a homozygous genotype contains two IBD alleles or not, we do not know precisely whether to count 1 or 2 distinct alleles so, in the numerator of eqn (12), we take a weighted average between 1 and 2, weighted by the probability that the two alleles are IBD. That gives us the (expected) total number of distinct A_k alleles in the dataset based on our current estimates for the X_{ij} values. In the denominator we count the total (expected) number of distinct alleles at marker j in the dataset. The updated estimate for p_{jk} is the estimated proportion of distinct alleles in the dataset that are of type A_{jk}. A more formal derivation of this algorithm is presented in the online Supplementary Material at http://journals.cambridge.org/GRH.

(vi) *New Method 3: an EM algorithm for joint estimation of inbreeding coefficients and allele frequencies including null alleles*

For this algorithm, we generalize our Method 2 algorithm by adding the possibility of null alleles. This algorithm can also be viewed as a generalization of Kalinowski and Taper's method in which we allow their model to include the possibility of inbreeding. We take the convention that there are now n_j+1 possible alleles at marker j ($j=1, \ldots, m$), where A_{j1}, $\ldots A_{jn_j}$ are the visible alleles and $A_{j,n}$ is the null allele. We adopt Kalinowski and Taper's model for missing data: that each genotype at marker j ($j=1, 2 \ldots, m$) has a probability β_j of being missing, independent of the genotype at that marker.

The presence of null alleles, as well as the fact that a genotype may be randomly missing, may cause a difference between the observed and true genotypes. Define g_{ij}^* to be individual i's true genotype at marker j and g_{ij} to be the observed genotype, with the convention that $g_{ij}=M$ denotes a missing genotype. Define Θ to be the values of all the parameters, so $\Theta=(\vec{f},\vec{p},\vec{\beta})$. As before, let X_{ij} be 0 or 1 depending on whether individual i's 2 alleles at marker j are IBD. Define B_{ij} ($i=1, \ldots, N$, $j=1, \ldots, m$) such that $B_{ij}=1$ if individual i's genotype at marker j is missing at random and $B_{ij}=0$ otherwise. The variable B_{ij} is unobserved because a missing genotype may be missing randomly or missing because the genotype is homozygous for a null allele.

The equations for the EM algorithm depend on two critical quantities, $\Pr(g_{ij}^*|g_{ij}, X_{ij}, \Theta)$, and $E[X_{ij}|g_{ij}, \Theta]$, $i=1, \ldots, N$, $j=1, \ldots, m$. Values for the first of these can be found in Table 1. Note that there are no formulae given for $\Pr(g_{ij}^*|g_{ij}=A_kA_l, X_{ij}=1, \Theta)$ in Table 1, since, if $X_{ij}=1$ (and so the 2 alleles are IBD), it is impossible to observe a heterozygous genotype under this model. The values for $E[X_{ij}|g_{ij}, \Theta]$ are given in the following equation.

$$E[X_{ij}|g_{ij},\Theta]=\begin{cases}\frac{f_ip_{jk}}{f_ip_{jk}+(1-f_i)(p_{jk}^2+2p_{jn}p_{jk})}, & g_{ij}=A_{jk}A_{jk},\\ 0, & g_{ij}=A_{jk}A_{jl},\ (k\neq l),\\ \frac{f_i(\beta_j+(1-\beta_j)p_{jn})}{\beta_j+(1-\beta_j)\left[fp_{jn}+(1-f)p_{jn}^2\right]}, & g_{ij}=M.\end{cases} \tag{13}$$

The likelihood is now

$$L(\Theta)=\prod_{i=1}^{N}\prod_{j=1}^{m}\Pr(g_{ij}|\Theta). \tag{14}$$

Table 1. *Conditional genotype probabilities of the form* $\Pr(g_{ij}^*|g_{ij}, X_{ij}, \Theta)$ *for each combination of* g_{ij}, g_{ij}^* *and* X_{ij}

		True genotype g_{ij}^*			
g_{ij}	X_{ij}	A_kA_k	$A_kA_l (k \neq l)$	A_kA_n	A_nA_n
A_kA_k	0	$\frac{p_{jk}}{p_{jk}+2p_{jn}}$	0	$\frac{2p_{jn}}{p_{jk}+2p_{jn}}$	0
A_kA_k	1	1	0	0	0
A_kA_l	0	0	1	0	0
M	0	$\frac{\beta_j p_{jk}^2}{\beta_j+(1-\beta_j)p_{jn}^2}$	$\frac{2\beta_j p_{jk}p_{jl}}{\beta_j+(1-\beta_j)p_{jn}^2}$	$\frac{2\beta_j p_{jk}p_{jn}}{\beta_j+(1-\beta_j)p_{jn}^2}$	$\frac{p_{jn}^2}{\beta_j+(1-\beta_j)p_{jn}^2}$
M	1	$\frac{\beta_j p_{jk}}{\beta_j+(1-\beta_j)p_{jn}}$	0	0	$\frac{p_{jn}}{\beta_j+(1-\beta_j)p_{jn}}$

Given the estimated parameter values after the zth iteration, namely $\Theta^{(z)}=(\vec{f}^{(z)}, \vec{p}^{(z)}, \vec{\beta}^{(z)})$, the EM algorithm proceeds as follows (as before, using the notation $E[X_{ij}|g_{ij}, \Theta^{(z)}]=E_z[X_{ij}]$ when convenient):

1. Update the values of $E[X_{ij}|g_{ij}, \Theta^{(z)}]$ for all i and j using eqn (13). Call these updated values $E_z[X_{ij}]$.
2. The updated estimates for the inbreeding coefficients are:

$$f_i^{(z+1)}=\frac{\sum_{j=1}^{m} E_z[X_{ij}]}{m}. \tag{15}$$

3. Update the allele frequency estimates using the following equations:

$$\begin{aligned} p_{jk}^{(z+1)}=\sum_{i=1}^{N}\big[&2P\left(g_{ij}^*=A_kA_k|g_{ij}, X_{ij}=0, \Theta^{(z)}\right)\\ &\times\left(1-E_z[X_{ij}]\right)+P\left(g_{ij}^*=A_kA_k|g_{ij}, X_{ij}=1, \Theta^{(z)}\right)\\ &\times E_z[X_{ij}]+\sum_{l\neq k} P\left(g_{ij}^*=A_kA_l|g_{ij}, X_{ij}^*=0, \Theta^{(z)}\right)\\ &\times(1-E_z[X_{ij}])\big]/C_j^{(z+1)}, \end{aligned} \tag{16}$$

where

$$\begin{aligned} C_j^{(z+1)}=\sum_{i=1}^{N}\sum_{l=1}^{n_j}\big[&2P\left(g_{ij}^*=A_lA_l|g_{ij}, X_{ij}=0, \Theta^{(z)}\right)\\ &\times\left(1-E_z[X_{ij}]\right)+P\left(g_{ij}^*=A_lA_l|g_{ij}, X_{ij}=1, \Theta^{(z)}\right)\\ &\times E_z[X_{ij}]+2\sum_{h<l} P\left(g_{ij}^*=A_lA_h|g_{ij}, X_{ij}=0, \Theta^{(z)}\right)\\ &\times(1-E_z[X_{ij}])\big]. \end{aligned} \tag{17}$$

The values for quantities of the form $P(g_{ij}^*|g_{ij}, X_{ij}, \Theta)$ can be calculated according to Table 1.

4. Update the probability that a genotype will be randomly missing:

$$p_{miss}^{(z)}=\beta_j^{(z)}+\left(1-\beta_j^{(z)}\right)\left[f_i^{(z)}p_{j,n}^{(z)}+\left(1-f_i^{(z)}\right)\left(p_{j,n}^{(z)}\right)^2\right]. \tag{18}$$

5. The updated estimates for β_j ($j=1, \ldots, m$) are

$$\beta_j^{(z+1)}=\frac{\left(\sum_{i=1}^{N}\left(\beta_j^{(z)}/p_{\text{miss}}^{(z)}\right)\right)}{m}. \tag{19}$$

A detailed derivation of this algorithm can be found in the online Supplementary Material at http://journals.cambridge.org/GRH.

(vii) *Simulations without null alleles*

We performed a number of simulations in order to evaluate the performance of our estimators and compare them with other estimators. We varied the number of genotyped markers ($m=5$, 10, 20, 50 and 100), number of alleles per marker ($a=2$, 5 and 10) and number of individuals in the simulated datasets ($N=20$ and 100). In all cases the allele frequencies at a marker with m alleles were ($1c$, $2c$, …, mc), where $c=2/(m(m+1))$. The markers were simulated to be unlinked and not in linkage disequilibrium with each other. A simulated dataset consisted of equal numbers of individuals with the following inbreeding coefficients: $f=0{\cdot}00$, $0{\cdot}02$, $0{\cdot}05$, $0{\cdot}10$, $0{\cdot}20$, $0{\cdot}30$, $0{\cdot}4$, $0{\cdot}5$, $0{\cdot}7$ and $0{\cdot}9$. Once an individual was assigned an inbreeding coefficient, its genotypes were simulated independently at each locus according to the probabilities given in eqn (5). For each dataset, we recorded the estimated inbreeding coefficient for 10 individuals – one with each of the possible inbreeding coefficients. One thousand datasets were generated for each combination of m, a and N.

For these simulations, we compared the behaviour of our Method 1 MLE with that of the Simple estimator and Ritland's estimator (and versions of these adjusted to give estimated values of f that lie between 0 and 1). Since these methods assume known allele frequencies, we assumed the true allele frequencies (the allele frequencies used to generate the data) when

we used these algorithms to estimate the inbreeding coefficients.

Considering the case in which allele frequencies were not assumed to be known, we also looked at the behaviour of our Method 2 MLE and the version of Vogl's estimators that did not account for null alleles. These were also compared with our Method 1 MLE, Ritland's estimator, and the Simple estimator where these methods were given allele frequencies naively estimated from the data using simple allele-counting methods.

Two criteria for judging the performance of an estimator are the bias and the mean-square-error (MSE). The bias of an estimator is the average amount by which it overestimates or underestimates a parameter. Let $\hat{f}_i$ be the estimated inbreeding coefficient for the ith individual analysed of those 1000. Then our estimate for the bias is the average value by which the estimator missed the true value f,

$$\text{BIAS} = \frac{\sum_{i=1}^{1000}\left(\hat{f}_i - f\right)}{1000}. \tag{20}$$

Since, within each dataset, we recorded the estimated inbreeding coefficient for one individual with each true inbreeding coefficient ($f = 0{\cdot}00$, $0{\cdot}02$, $0{\cdot}05$, $0{\cdot}10$, $0{\cdot}20$, $0{\cdot}30$, $0{\cdot}4$, $0{\cdot}5$, $0{\cdot}7$ and $0{\cdot}9$), we could use eqn (20) to estimate the bias for each true inbreeding coefficient. We knew that the estimated inbreeding coefficients would be biased upward when f was quite small, and looking at bias separately for different degrees of inbreeding would allow us to see how this bias decreases with increasing sample size.

The MSE of an estimator is the average squared distance between the value of the estimator and the parameter it is supposed to estimate. An estimate for the MSE of the estimators was found by taking

$$\text{MSE} = \frac{\sum_{i=1}^{1000}(\hat{f}_i - f)^2}{1000}. \tag{21}$$

In our summaries, we reported the RMSE, which is the square root of the MSE.

(viii) *Simulations with null alleles*

Each simulated dataset had loci with null alleles. The null alleles had variable frequencies as follows: the first simulated marker had $p_{\text{null}} = 0$, the second had $p_{\text{null}} = 0{\cdot}1$, the third had $p_{\text{null}} = 0{\cdot}2$, and this pattern was repeated (i.e. every third marker had a frequency of 0, etc.). At each marker the frequencies of the non-null alleles were given by

$$p_{jk} = \frac{2k}{n_j(n_j+1)}(1 - p_{\text{null}}), \tag{22}$$

where p_{jk} is the allele frequency of the kth observable allele at marker j, p_{null} is the frequency of the null allele and n_j is the number of observable alleles. The rate at which a genotype would be missing at random was $\beta = 0{\cdot}05$ at all loci.

Other than the addition of null alleles and randomly missing data, these simulations were performed similarly to the simulations without null alleles with the following exceptions: (1) We did not simulate the case in which the number of (visible) alleles at a marker was $a = 2$, (2) 400 datasets were generated for each combination of m, a and N, (3) we used the version of Vogl's algorithm that allows for the presence of null alleles.

(ix) *Real datasets*

To evaluate the performance of the estimators on real data, we looked at two datasets: the Mouse dataset (Laurie *et al.*, 2007) contained 94 wild-caught mice (*M. musculus domesticus*) captured in Arizona. We chose to use 757 single nucleotide polymorphism (SNP) loci with an average inter-marker spacing of 2 Mb. At this spacing, we expect that there is no linkage disequilibrium between the loci. We chose this dataset because mice are generally known to practice inbreeding (Ihe *et al.*, 2006). There was some missing data in this dataset; however, all 94 mice were genotyped at at least 714 of these markers.

The Deer dataset (Shaw *et al.*, 2006) consisted of 740 white-tailed deer (*Odocoileus virginianus*) genotyped at 15 microsatellite loci. The number of observed alleles at each locus ranged from 5 to 17. This dataset was of interest to us because DeYoung *et al.* (2003) had found null alleles at a number of these loci in another population of white-tailed deer, so this was a dataset for which we expected both inbreeding and null alleles.

3. Results

(i) *Simulation results*

Our first simulations were designed to investigate the behaviour of the estimators that analyse data from a single individual and assume known allele frequencies. Tables showing the estimated bias and RMSE values for the MLE, Ritland, Simple, adjusted Ritland and adjusted Simple estimators are included in the online Supplementary Material at http://journals.cambridge.org/GRH. It should be noted that, because allele frequencies were the same for all loci in our simulations, the Simple estimator is identical to the unweighted version of the estimator given in Carothers *et al.* (2006) for these simulations. Although the MLE and adjusted Simple estimators gave quite different estimates of f in many instances, both estimators yielded nearly identical estimates for RMSE and bias for all sample sizes and all values of f that we examined. The adjusted Ritland estimator performed

Table 2. *RMSE, allele frequencies unknown, 10 possible alleles at each marker*

Number of markers	Estimator*	*f* 0·0	0·05	0·1	0·2	0·3
10	jMLE	0·079	0·095	0·115	0·153	0·178
	MLE	0·077	0·092	0·111	0·150	0·176
	V1	0·150	0·131	0·120	0·119	0·131
	V2	0·017	0·050	0·099	0·199	0·284
	aSim	0·081	0·094	0·114	0·148	0·172
	aRit	0·064	0·074	0·094	0·139	0·177
20	jMLE	0·045	0·069	0·094	0·125	0·133
	MLE	0·045	0·066	0·090	0·122	0·132
	V1	0·098	0·086	0·085	0·098	0·108
	V2	0·010	0·052	0·102	0·190	0·252
	aSim	0·049	0·069	0·092	0·121	0·131
	aRit	0·035	0·055	0·080	0·118	0·147
50	jMLE	0·024	0·048	0·070	0·085	0·089
	MLE	0·025	0·046	0·066	0·084	0·090
	V1	0·057	0·048	0·050	0·069	0·081
	V2	0·001	0·050	0·097	0·159	0·151
	aSim	0·025	0·048	0·067	0·084	0·088
	aRit	0·021	0·041	0·067	0·099	0·125
100	jMLE	0·011	0·041	0·057	0·062	0·064
	MLE	0·013	0·037	0·053	0·063	0·067
	V1	0·037	0·030	0·038	0·053	0·058
	V2	0·001	0·050	0·091	0·102	0·070
	aSim	0·013	0·039	0·054	0·060	0·063
	aRit	0·010	0·037	0·059	0·087	0·117

*The estimators are denoted as follows: jMLE is our Method 2 MLE (which jointly estimates allele frequencies and inbreeding coefficients), MLE is the Method 1 MLE, V1 is Vogl's Bayesian estimator (assuming a uniform prior), V2 is Vogl's Bayesian estimator (assuming the prior suggested in Vogl *et al*., 2002), aSim and aRit are the adjusted Simple and Ritland estimators, respectively. The MLE, aSim and aRit estimators all used allele frequencies estimated from the data by allele counting.

similarly to the others with the exception that it tended to have a larger RMSE when *f* was large.

We next turned our attention to the situation in which allele frequencies were not known and had to be estimated from the dataset. In this case, we compared our Method 2 MLE and Vogl's estimators. We also included the Method 1 MLE, as well as the adjusted Ritland and Simple estimators, with the provision that these estimators would use allele frequencies estimated by simple allele counting from the dataset. Table 2 shows the estimated RMSE values for these estimators for the situation in which there were 20 individuals in each simulated dataset. More results appear in the online Supplementary Materials. Even with just 20 individuals in the sample, these estimators performed similarly to the case in which allele frequencies were known. It did not seem to matter much whether allele frequencies were estimated using sophisticated methods (e.g. our Method 2 MLE or Vogl's methods) or by simple allele counting. Also apparent from the table is that Vogl's method that used a strongly informative prior (V2, which begins with a strong assumption that the values of *f* will be close to 0 or 1) gives an inflated RMSE compared with the other estimators when *f* is moderately large and the number of markers is not large. The other side of this phenomenon is that this estimator did very well when *f* was small, which is not surprising since the estimator is based on a strong prior assumption that *f* will be close to 0.

Next, we considered the case in which null alleles were present in the data. Table 3 shows the mean bias observed among 400 datasets for each of several inbreeding coefficients when the simulated datasets each consisted of 20 individuals. Tables showing RMSE are included in the online Supplementary Materials. Three things are apparent from Table 3: (1) the estimators that ignored the presence of null alleles (the adjusted Ritland and Simple estimators and the Method 2 MLE) tended to over-estimate inbreeding coefficients when null alleles are present. (2) The Method 3 MLE was able to dramatically reduce the bias compared with the estimators that ignore null alleles. (3) For the Vogl algorithm, neither prior used in our analyses was entirely satisfactory. Under the uniform prior, the Vogl algorithm systematically

Table 3. *Estimated bias in the presence of null alleles, 10 alleles per marker*

Number of loci	Estimator	*f* 0·0	0·05	0·1	0·2	0·3
10	jMLE	0·141	0·115	0·099	0·083	0·063
	aSim†	0·145	0·123	0·103	0·087	0·064
	aRit	0·096	0·055	0·030	−0·027	−0·069
	M3MLE	0·059	0·030	0·009*	−0·010*	−0·022*
	V1‡	0·204	0·170	0·138	0·090	0·044
	V2	0·000*	−0·048	−0·096	−0·196	−0·285
20	jMLE	0·127	0·108	0·106	0·096	0·080
	aSim	0·130	0·113	0·109	0·097	0·080
	aRit	0·086	0·053	0·034	−0·014	−0·064
	M3MLE	0·030	−0·004*	−0·007*	−0·019*	−0·026*
	V1	0·149	0·118	0·100	0·066	0·029
	V2	0·000*	−0·050	−0·094	−0·185	−0·267
50	jMLE	0·131	0·118	0·121	0·087	0·081
	aSim	0·133	0·120	0·120	0·088	0·080
	aRit	0·087	0·060	0·042	−0·020	−0·064
	M3MLE	0·012	−0·019*	−0·033*	−0·055*	−0·033*
	V1	0·100	0·074	0·062	0·019	0·011
	V2	0·001*	−0·047	−0·092	−0·163	−0·173
100	jMLE	0·135	0·122	0·113	0·090	0·076
	aSim	0·137	0·123	0·114	0·089	0·075
	aRit	0·089	0·061	0·035	−0·015	−0·067
	M3MLE	0·003	−0·032*	−0·051*	−0·059*	−0·044*
	V1	0·078	0·053	0·041	0·013	0·002
	V2	0·001*	−0·045	−0·075	−0·102	−0·077

*Indicates a value that is not significantly different than 0 at a significance level of $\alpha = 0{\cdot}05$.
†The adjusted Simple and Ritland algorithms used in these analyses used allele frequencies estimated for each dataset by allele counting.
‡The Vogl algorithms used in these analyses were the versions of the Vogl algorithm that account for null alleles.

over-estimated the degree of inbreeding for individuals that had low-inbreeding coefficients. The Vogl algorithm with the prior suggested in Vogl *et al*. (2002) resulted in inbreeding coefficients being substantially under-estimated for highly inbred individuals. In fact, this algorithm tended to assign inbreeding coefficients close to 0 even for individuals whose true inbreeding coefficients were as high as 0·5 when there were up to 50 markers genotyped (results not shown).

The Method 3 MLE, Vogl's method and the algorithm of Kalinowski and Taper were all constructed with the ability to estimate null allele frequencies. To examine each algorithm's ability to give accurate estimates of the null allele frequently, we looked at the estimated null allele frequency for just three markers from each simulated dataset: one marker for each of the three possible null allele frequencies that were present in the simulations. Table 4 shows the average estimated null allele frequency over 400 datasets when there were 20 individuals in each dataset and 10 observable alleles at each marker. As might be expected, the Kalinowski and Taper algorithm, which was developed for use on datasets with non-inbred individuals, tends to over-estimate the frequencies of null alleles considerably. The Vogl estimator that assumes the beta (0·001, 0·001) prior performs similarly. Both of these algorithms assume lower inbreeding coefficients than actually exist in the dataset: the Kalinowski and Taper algorithm assumes that $f = 0$ for all individuals and the Vogl algorithm with this prior begins with a strong assumption that the inbreeding coefficients will be close to 0 and so tends to produce under-estimates of f for individuals with larger inbreeding coefficients. When inbreeding coefficients are under-estimated, the excess homozygosity observed in the data results in inflated null allele frequencies. The Vogl estimator with the uniform prior and the Method 3 MLE, which do better at estimating the inbreeding coefficients, consequently give more accurate estimated null allele frequencies.

(ii) *Real data results*

For both the mouse and deer datasets, we compared the behaviour of the Method 3 MLE and the Vogl estimator that assumed a uniform prior. The top row of Fig. 1 shows some results from the analysis of the mouse data. For this dataset, the Vogl estimator nearly always gave considerably smaller estimated

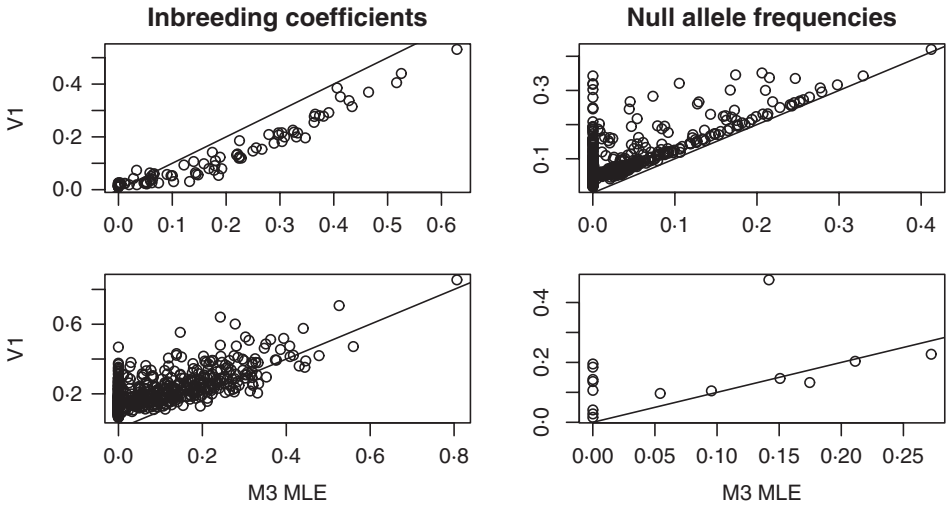

Fig. 1. Results from the analyses of the Arizona Mice (top row) and Deer (bottom row) datasets. The plots on the left contain a point for each individual (mouse or deer) in the dataset. The *x*- and *y*-coordinates of each point are the estimated inbreeding coefficients for that individual from the Method 3 MLE (M3) and the Vogl estimator with the uniform prior (V1), respectively. In the plots on the right, each point represents a marker in the dataset and the coordinates are the estimated frequencies of the null allele at that marker.

Table 4. *Mean estimated null allele frequencies, 10 alleles per locus*

Number of loci	Estimator	Null allele frequency 0·0	0·1	0·2
10	M3MLE	0·013	0·075	0·152
	V1	0·070	0·115	0·167
	V2	0·135	0·200	0·260
	K*	0·154	0·231	0·305
20	M3MLE	0·017	0·082	0·159
	V1	0·073	0·121	0·176
	V2	0·122	0·185	0·246
	K	0·159	0·233	0·302
50	M3MLE	0·022	0·084	0·175
	V1	0·074	0·117	0·185
	V2	0·099	0·152	0·224
	K	0·160	0·226	0·304
100	M3MLE	0·022	0·093	0·176
	V1	0·074	0·122	0·184
	V2	0·090	0·143	0·208
	K	0·155	0·234	0·303

*K denotes the Kalinowski and Taper estimator.

inbreeding coefficients than the MLE. Consequently (because excess homozygosity must be due to either null alleles or inbreeding), Vogl gave higher estimates for the frequency of null alleles than did the MLE. Interestingly, there were many markers for which the Vogl and MLE estimators were in sharp disagreement regarding the frequency of null alleles. Specifically, there were 148 (out of the 757) markers for which the MLE gave null allele frequencies below 0·01 and the Vogl estimator gave null allele frequencies above 0·05. This type of behaviour can be explained as follows: in the version of Vogl's estimator that we used (the version explicitly described in Vogl *et al*., 2002), the model assumes that any missing genotype is homozygous for a null allele. Hence, Vogl's algorithm will assign a high null allele frequency to any marker with a substantial number of missing genotypes. The model used in calculating our MLE, however, following Kalinowski & Taper (2006), has a parameter that allows genotypes to be missing at random at a marker. With this parameter in the model, the MLE will not assign a high null allele frequency to markers that have missing data but no other sign of excess homozygosity beyond that accounted for by the inbreeding coefficients of the individuals in the dataset.

The second row of Fig. 1 shows some results from the analysis of the deer data. As with the mouse data, we focused on the behaviour of the Method 3 MLE and the Vogl estimator that assumed a uniform prior. The estimators are in rough agreement about the estimated inbreeding coefficients for each deer in the dataset, with the MLE tending to give lower estimates than Vogl's estimator. When estimating null allele frequencies, the two estimators tended to give similar results except for markers with a high degree of missing data, in which case the Vogl estimator yielded much higher estimates.

4. Discussion

In our analyses, we found that the MLE algorithms presented in the paper work generally well. We found weaknesses in most of the other estimators we examined. For example, Ritland's estimator had problems when the true inbreeding coefficients were large

and Vogl's estimator was sensitive to the choice of parameters of the prior distribution and tended to over-estimate null allele frequencies when confronted with missing data. As pointed out in Vogel *et al.* (2002), however, Vogl's algorithm could be easily adjusted to handle missing data more effectively. The Simple estimator, which is similar to Carother's estimator, was the only estimator examined that seemed to perform as well as the MLE. Unfortunately, it is not obvious how to adjust the Simple estimator to account for null alleles.

The bias of the MLE is virtually identical to the bias of the adjusted versions of the moment estimators (see the online Supplementary Materials). The conclusion to be drawn is that the bias of the MLE is completely inconsequential for most purposes as it seems to be primarily the result of assigning an inbreeding coefficient of 0 rather than a negative value to individuals that show less homozygosity than expected. There is one situation in which the moment estimators might be preferred to the MLE, namely, if a researcher is interested in showing that individuals in a population are showing less heterozygosity than would occur by random mating (e.g. this might occur if the population is showing an avoidance of matings between relatives). The moment estimators we looked at are capable of giving the negative inbreeding coefficients that could indicate that sort of scenario. As the MLEs are based on a model that views f as a probability rather than a correlation coefficient, they cannot give negative values.

In order to investigate the robustness of the estimators to various strange situations or violations of model assumptions, we performed additional simulations, the details of which are included in the online Supplementary Materials. The first set of these simulations looked at the performance of the estimators on datasets consisting of sets of closely related individuals. In our simulations, all estimators performed poorly under these circumstances. The second question was whether the estimators that allow for null alleles would have trouble if there were no null alleles or if all markers had the same null allele frequencies. In our simulations, this seemed to pose no difficulty. Finally, the third set of additional simulations was designed to see the effect of having certain individuals for which many genotypes were missing due to degradation. This common situation is different from that modelled by the Method 3 MLE (in which genotypes are missing at random for all individuals, or are missing because they are homozygous for null alleles) or by Vogl's algorithms (which assume missing genotypes are homozygous for null alleles). In our simulations, this type of missingness was disastrous. All three estimators greatly overestimated the inbreeding coefficients for these individuals. In our simulations, these individuals consisted of just 10 out of the 100 individuals in the simulated datasets. Their inclusion did not seem to have a great effect on the estimated inbreeding coefficients of the other individuals in the datasets.

An estimator that was not included in this study is that of Wang (2011*a*). Wang's estimator is a Bayesian estimator that accounts for null alleles and allelic dropout. Wang uses an exponential distribution as the prior distribution for inbreeding coefficients. In Wang's paper, the likelihood equation was written out for a single individual. Allele frequencies (including the null allele frequency) and the dropout rate were required to be known values. As written, then, Wang's estimator is not directly comparable to any of our MLEs. To make our Method 3 estimator comparable to Wang's estimator would involve treating the allele frequencies and error rates in our model as known constants rather than values that are estimated using our EM algorithm. Similarly, Wang's method could be modified to estimate null allele frequencies and/or error rates as follows: the likelihood for a set of individuals would be the product of Wang's likelihood for single individual. Using this new likelihood function (and, presumably, a prior distribution for the inbreeding coefficients that would be the product of exponential distributions), the allele frequencies and dropout rates could be estimated along with the inbreeding coefficients. The difference in which parameters are considered known and which are estimated constitutes a minor difference between our Method 3 estimator and Wang's estimator. Apart from this, there are three actual differences between the estimators: (1) Wang begins by assuming a prior distribution for the inbreeding coefficient that will tend to gently push his estimated inbreeding coefficients towards zero, (2) Wang's model includes the possibility of allelic dropout (but only for genotypes that are heterozygous for non-null alleles), which is a source of excess homozygosity that is ignored in our model and (3) Wang's model uses only non-missing data, and hence, does not make use of the fact that, when null alleles exist, missing data itself provide some evidence of homozygosity.

A key point that is apparent from our simulations is that the variability of all the estimators we examined will be substantial unless a large number of markers is used (see the RMSE values in the tables in the online Supplementary Materials). For example, when there are unknown allele frequencies and no null alleles, and 50 markers have been genotyped, each with 10 alleles, the Method 2 MLE yields a RMSE of 0·048 for individuals with $f=0{\cdot}05$. Since the bias is small, the RMSE will be close to the standard deviation. A margin of error for an estimator will be roughly twice its standard deviation, so, in this case, even with 50 markers, the margin of error for the estimate will be close to 0·10. Hence, even with 50 highly variable

markers, the estimated inbreeding coefficients will be quite imprecise.

Researchers intending to use an estimated inbreeding coefficient in a subsequent linear regression against, for example, a fitness-related trait, should be aware of how uncertainty in estimates of f may affect their results. There is a body of literature on the effects of measurement error on regression analyses (e.g. Frost & Thompson, 2000) and it can be shown that variability in the predictor variable (estimated inbreeding coefficients) will result in a systematic tendency to under-estimate the strength of the relationship between the predictor and response variables (inbreeding and fitness). This will reduce the power of studies looking for evidence of inbreeding depression.

This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Declaration of interests

None.

References

Anderson, A. & Weir, B. (2007). A maximum-likelihood method for the estimation of pairwise relatedness in structured populations. *Genetics* **176**, 421–440.

Carothers, A. D., Ruden, I., Kolcic, I., Polasek, O., Hayward, C., Wright, A. F., Campbell, H., Teague, P., Hastie, N. D. & Weber, J. L. (2006). Estimating human inbreeding coefficients: comparison of genealogical and marker heterozygosity approaches. *Annals of Human Genetics* **70**, 666–676.

Chapuis, M.-P. & Estoup, A. (2007). Microsatellite null alleles and estimation of population differentiation. *Molecular Biology and Evolution* **24**, 621–631.

Chybicki, I. J. & Burczyk, J. (2009). Simultaneous estimation of null alleles and inbreeding coefficients. *Journal of Heredity* **100**, 106–113.

DeYoung, R. W., Demarais, S., Honeycutt, R. L., Gonzales, R. A., Gee, K. L. & Anderson, J. D. (2003). Evaluation of a DNA microsatellite panel useful for genetic exclusion studies in white-tailed deer. *Wildlife Society Bulletin* **31**, 220–232.

Frost, C. & Thompson, S. G. (2000). Correcting for regression dilution bias: comparison of methods for a single predictor variable. *Journal of the Royal Statistical Society A* **163**, 173–190.

Girard, P. (2011). A robust statistical method to detect null alleles in microsatellite and SNP datasets in both panmictic and inbred populations. *Statistical Applications in Genetics and Molecular Biology* **10**, Article 9 [online].

Ihe, S., Ravaoarimanana, I., Thomas, M. & Tautz, D. (2006). An analysis of signatures of selective sweeps in natural populations of the house mouse. *Molecular Biology and Evolution* **23**, 790–797.

Kalinowski, S. T. & Taper, M. L. (2006). Maximum likelihood estimation of the frequency of null alleles at microsatellite loci. *Conservation Genetics* **7**, 991–995.

Laurie, C. C., Nickerson, D. A., Anderson, A. D., Weir, B. S., Livingston, R. J., Dean, M. D., Smith, K. L., Schadt, E. E. & Nachman, M. (2007). Linkage disequilibrium in wild mice. *PLoS Genetics* **3**, 1487–1495.

Leutenegger, A.-L., Prum, B., Génin, E., Verny, D., Lemainque, A., Clerget-Darpoux, F. & Thompson, E. A. (2003). Estimation of the inbreeding coefficient throught use of genomic data. *American Journal of Human Genetics* **73**, 516–523.

Li, C. C. & Horvitz, D. G. (1953). Some methods of estimating the inbreeding coefficient. *American Journal of Human Genetics* **5**, 107–117.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M., Bender, D., Maller, J., Sklar, P., de Bakker, P., Daly, M. & Sham, P. (2007). Plink: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics* **81**, 559–575.

Ritland, K. (1996). Estimators for pairwise relatedness and inbreeding coefficients. *Genetical Research* **67**, 175–186.

Shaw, J. C., Lancia, R. A., Conner, M. C. & Rosenberry, C. S. (2006). Effect of population demographics and social pressures on white-tailed deer dispersal ecology. *Journal of Wildife Management* **70**, 1293–1301.

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

Vogl, C., Karhu, A., Moran, G. & Savolainen, O. (2002). High resolution analysis of mating systems: inbreeding in natural populations of *Pinus radiata*. *Journal of Evolutionary Biology* **15**, 433–439.

Wang, J. (2011*a*). A new likelihood estimator and its comparison with moment estimators of individual genome-wide diversity. *Heredity* **107**, 433–443.

Wang, J. (2011*b*). Unbiased relatedness estimation in structured populations. *Genetics* **187**, 887–901.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics* **11**, 95–103.

Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *American Journal of Human Genetics* **88**, 76–82.