Molecular Ecology (2006) **15**, 1657–1667 doi: 10.1111/j.1365-294X.2006.02873.x

A new method to estimate relatedness from molecular markers

J. FERNÁNDEZ and M. A. TORO
Departamento de Mejora Genética Animal. Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria, Ctra. Coruña Km. 7,5, 28040 Madrid (Spain)

Abstract

Four major problems can affect the efficiency of methods developed to estimate relatedness between individuals from information of molecular markers: (i) some of them are dependent on the knowledge of the true allelic frequencies in the base population; (ii) they assume that all loci are unlinked and in Hardy–Weinberg and linkage equilibrium; (iii) pairwise methods can lead to incongruous assignations because they take into account only two individuals at a time; (iv) most are usually constructed for particular structured populations (only consider a few relationship classes, e.g. full-sibs vs. unrelated). We have developed a new approach to estimate relatedness that is free from the above limitations. The method uses a 'blind search algorithm' (actually *simulated annealing*) to find the genealogy that yield a co-ancestry matrix with the highest correlation with the molecular co-ancestry matrix calculated using the markers. Thus (i and ii) it makes no direct assumptions about allelic frequencies or Hardy–Weinberg and linkage equilibrium; (iii) it always provide congruent relationships, as it considers all individuals at a time; (iv) degrees of relatedness can be as complex as desired just increasing the 'depth' (i.e. number of generations) of the proposed genealogies. Computer simulations have shown that the accuracy and robustness against genotyping errors of this new approach is comparable to that of other proposed methods in those particular situations they were developed for, but it is more flexible and can cope with more complex situations.

Keywords: co-ancestry, genealogy reconstruction, molecular markers, simulated annealing

Received 16 September 2005; revision accepted 8 December 2005

Introduction

Knowledge of the pedigree structure of a population, or the co-ancestries between the individuals which belong to it, is important in many areas of the population biology and genetics. For example, it is needed to estimate genetic parameters (heritabilities and genetic correlations) and breeding values (Lynch & Walsh 1998); it helps to avoid the loss of diversity and the increase of inbreeding in the management of populations (Caballero & Toro 2000, 2002); it allows for the study of mating systems, parental care, dispersal and other parameters in ecological or behavioural genetics (Avise 1994; Frankham *et al*. 2002). However, most natural populations and many captive ones lack this genealogical information, and it has been suggested that it should be possible to infer genealogies from the molecular data.

Following Butler *et al*. (2004), approaches developed in that field can be classified into two general groups: methods that generate complete population structures, and therefore involve explicit pedigree reconstruction, and pairwise methods that do not imply such pedigree reconstruction. The latter group can be further subdivided into methods that estimate relatedness for pairs of individuals based on their genotypic similarity for markers (see reviews by Toro *et al*. 2002 and Milligan 2003) and likelihood techniques allowing classifying pairs of individuals into different classes of relationships based on marker information (Herbinger *et al*. 1997; Mousseau *et al*. 1998).

The nature of each estimator and the assumptions made when designing them may lead to different problems that are summarized below. Obviously, not

Correspondence: Jesús Fernández Martín, Fax: 34-913572293; E-mail: jmj@inia.es

all methods suffer all the shortcomings. For pairwise methods, the first concern is that they can lead to incongruous assignations because only two individuals are taken into account at a time and therefore the resulting co-ancestry matrix could be non-positive definite. For example, individuals A and B can be classified as full-sibs, B and C classified as full-sibs too, but estimated relationship between A and C might be other than full-sibs. In the opposite side, the exclusion of three homozygotes for different alleles being full-sibs cannot be done in a pairwise comparison basis (Thomas & Hill 2000).

Another problem linked to pairwise methods is that they are dependent on the knowledge of the allele frequencies of the assumed base population. The estimation of those frequencies from the population under study leads to different bias due to the existence of relatives in the sample (Ritland 1996) and the probable loss of alleles from the base (founder) population we use as reference (Toro *et al*. 2002). A final problem of pairwise methods is that, if we know that only certain types of relations exist (e.g. full-sibs, half-sibs or unrelated) we have to define arbitrary thresholds to decide the category to which a particular pair is assigned.

Several methods involving explicit pedigree reconstruction have been developed and have been recently reviewed by Butler *et al*. (2004). Some of them use MCMC algorithms to locate a partition that maximizes the full joint likelihood of the proposed family configuration (Thomas & Hill 2000, 2002; Emery *et al*. 2001; Smith *et al*. 2001; Wang 2004) or that maximizes a configuration score derived from pairwise likelihood ratio (Smith *et al*. 2001). Almudevar & Field (1999) proposed a method based on the exclusion principle that looks for the largest feasible full-sibs families, using partial likelihoods to choose between families of the same size. Another algorithm found in the literature is based on the Simpson's index of concentration (Butler *et al*. 2004), where marker information is only used to make sure that groups conform to Mendelian inheritance rules. Although the latter methods avoid the problems of pairwise estimators, they are still dependent on the knowledge of the true allelic frequencies (to calculate likelihoods) or they are very complex and computationally demanding and therefore they have been usually restricted to the estimation of full-sibs families or nested half-sibs designs (but see Emery *et al*. 2001 and Wang 2004). As an extension of the matter we dealt with in this study, Almudevar (2003) proposed an algorithm that use *simulated annealing* to reconstruct pedigrees when individuals from more than one generation where available (genotyped).

A common characteristic of almost all methods is that they assume that molecular data are error free. Butler *et al*. (2004) tested the robustness of different estimators against errors/mutations in the data concluding that, in general, biases were very large for all methods when a 5% of genotypes were incorrect. Recently, Wang (2004) have proposed a method that accounts for the possibility of errors when calculating the likelihood of each family configuration.

Toro *et al*. (2002) compared eight pairwise estimators of co-ancestry on data from 62 pigs genotyped for 49 microsatellites with the genealogical values calculated from the complete genealogy going 20 generations back. The main finding of their work was that the behaviour of all of them was similar, giving underestimates of the genealogical values. Simulation results indicate that the bias is due to the lack of information on the frequencies of the base (founder) population. However, the correlations between true and estimated co-ancestries were usually moderate or high, and similar to the correlations between genealogical co-ancestry and simple measures of molecular resemblance such a molecular co-ancestry. Molecular co-ancestry is defined as the probability that two alleles taken at random in the same locus, one from each individual, are equal (average across loci). This parameter is similar to the genealogical co-ancestry coefficient (Malécot 1948) but it measures *identity by state* instead of *identity by descent*.

The objective of the present study is to propose a simple and general method to reconstruct genealogies (or estimate co-ancestries) from molecular marker data, based on the searching for the genealogy which generates the pedigree co-ancestry matrix with the highest correlation with the molecular co-ancestry matrix. Due to its nature, the present algorithm is free from the problems and limitations previously pointed out for other methods. Computer simulations have been carried out to test for the accuracy and robustness against errors in the data of the method under different scenarios. It also has been applied to a real data set.

Materials and methods

Simulated data set

Two types of simulated data sets were used.

Hierarchical design. A number of males (M) and females ($2M$) were generated, randomly drawing alleles from a uniform distribution independently for each marker locus. Therefore, all allelic frequencies were equal, genotypic frequencies were in Hardy–Weinberg equilibrium and linkage equilibrium existed between loci. Afterwards, each male was mated to two females in a hierarchical design and five offspring were generated from each couple following Mendelian rules and assuming no linkage. Values for M were 5, 10, 20, 30 and 40 and therefore the number of individuals to estimate their relationship from (N) was 50, 100, 200, 300 and 400, respectively. The number of alleles for all markers was five. Different levels of molecular information were considered by simulating 5, 10 or 20 marker loci. This design is similar to the one described in Thomas & Hill (2002).

Unbalanced design. The same number of males and females (M) were generated as explained in the previous model. Then, $N = 100$ offspring were generated from randomly chosen parents. This way, not only sire half-sib families could appear but also dam half-sib families. This particular procedure also made the family size to be not equal (neither within replicates nor between replicates within the same set of parameters). Values for M were 5, 20 and 50 to simulate different degrees of relationship in the population. The number of simulated markers was 5, 10 and 20 and the number of equally frequent alleles was 2 (only in the case of 20 markers), 5 or 10 per locus. For scenarios with five markers with five alleles and 20 markers with two alleles, some data sets were generated using 'deeper' genealogies: two or four generations of size 2 M (half of each sex) of random mating were simulated prior to the generation of the 100 genotyped offspring.

Genotyping errors

For the unbalanced design, some simulations were run including errors in the genotype of individuals. To do so, 1% or 5% of the alleles of the population were randomly changed to another possible allele from the same locus. This case corresponds to the situation where mutations or unspecific human errors have occurred (class II errors in Wang 2004).

Real data set

The new method was also tested on real data from 62 pigs genotyped for 49 microsatellites (see Toro *et al*. 2002 for detailed characteristics of the population). Genealogy for that population was available for up to 20 generations back and therefore estimated co-ancestries could be compared with genealogical values. Three cases were considered: (i) using all available molecular information; (ii) using one marker per chromosome (data on a total of 18 microsatellites); (iii) using information from 10 randomly chosen markers.

Algorithm

The rationale behind the proposed method (FT thereafter) was looking for a feasible genealogy such that the co-ancestry matrix calculated from that pedigree had the highest correlation with the molecular co-ancestry calculated from the genotype of the target individuals for a set of molecular markers.

Molecular co-ancestry was calculated as the probability of two alleles, one from each individual, taken at random in the same locus being equal (and averaged across loci). For each generated genealogy, the pedigree-based co-ancestry was calculated following the classical rules (Falconer & Mackay 1996). Pearson's correlation between matrices was calculated for the genealogical and molecular co-ancestry of every pair of genotyped individuals. Self-co-ancestries were not included in the calculation of the correlation between matrices because they are dependent on the inbreeding level of an individual. Thus, for one-generation genealogies (where all individuals are not inbred and all pedigree self-co-ancestries are equal) more accuracy is to be found when not accounting for the values in the diagonal. For genealogies comprising more generations values were quite similar, whether we included self-co-ancestries or not. Only elements above the diagonal were used in the calculation, because both matrices (molecular and genealogical) are symmetric and therefore correlation was equal than if we used reciprocal relationships (and in return speed was increased).

We used a *simulated annealing* algorithm (Kirpatrick *et al*. 1983) to search across the sets of feasible family structures. The method started from a random solution, obtained by assigning a 'virtual' father and a 'virtual' mother (taken from a predefined number of available males and females) to every individual in the population. When there was no prior information, the number of potential parents was set to its maximum value, i.e. twice the number of individuals in the target population (half of each sex). If a more complex structure than full- and half-sibs families was to be considered, random grandparents, great-grandparents and so on were assigned for the number of generations allowed. In the latter situation, the same number of males and females was assumed to be available in all generations. The maximum number of generations considered in the present study was four. Notice that none of the ancestors of the target population was assigned a genotype and they are simply used as links between the members of the population.

The initial solution was tested to be compatible with Mendelian inheritance for the last generation and therefore every full-sib family in the target population conformed to the following rules: (i) no more than four different alleles and genotypes could exist in a family; (ii) a particular allele could be in heterozygosity with two other alleles at most; (iii) if there were four alleles, no homozygotes could be found; (iv) if there were three alleles, only one type of homozygote could exist. No tests were performed to check for the congruency of other type of relatives (e.g. half-sibs families) because the number of possible configurations is huge and the probability of detecting an incompatibility is very low.

From this starting point, alternative solutions were generated by substituting one of the ancestors of a random individual in a particular generation by another ancestor also chosen at random. Before further consideration, if the changed ancestor was the father or the mother of the individual (i.e. it belonged to the last but one generation), new solutions were tested to be congruous with Mendelian

inheritance, as we did with the initial solution. If the change led to the formation of an incompatible full-sib family, another random parent was drawn until compatibility was fitted. Then the 'values' of the present and the alternative solution (i.e. the correlation between the molecular co-ancestry matrix and the genealogical co-ancestry calculated from, respectively, the two considered pedigrees) were calculated. Due to its nature, *simulated annealing* is a minimization algorithm but we wanted to maximize the correlation. Therefore, we had to change the sign of both values in order to find the proper optimum. Acceptance of the alternative solution occurred with a probability $\Omega = \exp(-\Delta/T)$ where Δ was the difference between values of the alternative and actual solutions and T was a 'cooling' factor or temperature. Obviously, if $\Delta < 0$, $\Omega > 1$ and the alternative solution was always accepted because it was better. Five thousand alternative solutions were generated and tested and, then, the value of T was reduced by a factor Z. Another 5000 solutions were generated, the parameter T reduced and so on. Up to 150 steps (i.e. different values for T) were allowed. At the beginning of the process, many alternative solutions were accepted but as T decreased, it became more difficult to accept new solutions, unless they were better than the present one. The optimal solution was assumed to be found when no alternative solution was accepted for 5000 modifications generated at a given 'temperature', since the solution was very unlikely to be improved upon at either the current or lower temperatures, or when the maximum number of steps were performed.

In the first step, alternative solutions were constructed by changing ancestors for a large number of individuals (actually 10) at a time in order to perform a broad search across the space of solutions. But as the process advanced, smaller modifications were produced to allow for a fine searching in a particular area. The number of ancestors changed per solution was automatically decided by the algorithm and was dependent on the number of rejections/acceptances along the previous step. Actually, this number was 10 multiplied by the proportion of alternative solutions accepted in the previous step.

The rate of decrease of the cooling factor T (Z) was set to 0.9 based on previous simulations performed to 'tune' the algorithm. Too rapid decrease of T leads the algorithm to get stack in nonoptimal solutions, while too slow reduction is a waste of computing time. The training period for the algorithm allowed for finding an equilibrium value.

Measures of accuracy

The first measure of the good-of-fit between the true and estimated genealogies we used was based on the proportion of correctly and incorrectly reconstructed relationships (Smith *et al*. 2001; Thomas & Hill 2002). When only considering individuals to be full-sibs (FS), half-sibs (HS) or nonrelated (NR), there are nine situations depending on the true relationship and the relationship we estimated. Three of them correspond to correctly estimated relationships, another three are overestimations (presumed FS when really HS or NR and estimated as HS when NR) and the remaining three are underestimations (the opposite situations). The advantage of such a measure is that it allows specifying the direction (i.e. is not equivalent classifying an FS pair as NR than estimating NR couples as FS) and magnitude of errors (i.e. is not equivalent classifying an FS pair as HS or NR). However, as pointed out by Butler *et al*. (2004), these proportions could be less informative in extreme family configurations where there are a very small number of relatives. In such situations, failing a single relationship might cause proportions to fall a large amount. To avoid this problem, for some cases, absolute number of assigned pairs instead of proportions will be presented. This measure also loses meaning when more than single-generation pedigrees were constructed because the number of type of relationships increases rapidly.

Another measure of the bias of the estimator was the root mean square error (RMSE; Milligan 2003), and it was quantified as

$$\sqrt{\frac{\sum_{i=1}^{P}(\hat{\theta}_i - \theta_i)^2}{P}},$$

where $\hat{\theta}_i$ and θ_i were the estimated and the true co-ancestry, respectively, between the i^{th} couple, and P is the total number of possible comparisons (i.e. couples) which equals $N * (N-1)/2$.

The third measure of accuracy was the correlation between true and estimated co-ancestry matrices, calculated as well for all possible pairs of individuals.

Confidence on the estimations

When dealing with empirical data we have no idea of the true relationships but still would like to have a measure of the confidence on the estimation. Ideally, we would like to be able to calculate such a measure not for the entire population but for relationships between particular couples. One possible solution is to perform a bootstrap over loci for the molecular data. This way, we would obtain a confidence interval to test the reliability of each of the assignations. This procedure is highly computing demanding and would not be suitable for simulations but could be used in real data sets.

In order to get a faster measure of confidence, we proposed the following procedure. Not only the best but the 1000 best solutions were stored. Then, the co-ancestry between pairs of individuals was calculated for all kept

solutions and the mean and variance of each relationship computed across solutions. The variance of the estimations gave us a measure of the variability of each inferred relationship. Small variances (ideally null variances) meant that co-ancestry was most (all) of the times the same for that particular couple in the set of best solutions. Therefore, we could be quite confident on that assignation. If the value for the best solution (the point estimation) was far from the average value for the set of solutions, it posed some doubt on the correctness of the assignation.

In the case of single-generation genealogies, the proposed procedure was somehow equivalent to counting the number of kept solutions where the couple was assigned the same relationship (i.e. number of times classified as FS, HS or NR) in the same way that Emery *et al*. (2001) describe for its MCMC method.

Results

Simulated data

Figure 1 shows results from simulations involving hierarchical data. Left panels present the proportion of right or wrong assignments when pairs are actually full-sibs while right panels shows proportions when pairs are half-sibs. Lines with circles represent the proportion of pairs estimated as being full-sibs (FS), lines with diamonds those estimated as half-sibs (HS), and lines with crosses the ones estimated as nonrelated (NR). Solid lines are results from the present study while dashed lines are the results from Thomas & Hill (2002).

As pointed out by the authors, Thomas & Hill method (TH) was very sensitive to the level of marker information and the population size, producing large underestimates when few markers were genotyped, especially for large populations (top panels). Contrarily, the new method (FT), although it followed the same trends than TH method, was less affected by the population size, because the decrease in accuracy as the number of individuals increase was not so steep with our method. This can be clearly seen in HS couples assignation (right panels) where TH method performed better for small populations and 10 or 20 markers genotyped, but produced more errors than FT for populations with more than 20 families. New method was also advantageous in scenarios with little information (i.e. five markers genotyped), yielding less errors than TH for HS assignation and smaller errors for FS assignation.

Table 1 shows the distribution of correct and wrong assignments (and their standard errors) produced by FT method for unbalanced populations and different combinations of genotyped marker loci, number of alleles per locus and number of parents generating the problem population. Absolute values instead of proportions are shown to make fair comparisons between populations with different degree of global co-ancestry (i.e. arising from a different number of parents).

In general, agreement between estimated and true relations was quite good, with the number of correct assignations ranging from 70% (5 markers with 5 alleles in a population coming from 10 parents) to 98% (20 markers of 10 alleles and 100 parents). The direction of errors was dependent on the degree of global co-ancestry. For populations highly related the method tended to produce underestimations while for those loosely related there was a higher proportion of upward errors.

Correlations between true and estimated co-ancestry matrix as well as the root mean square error (in italics) from all pairs are shown in Table 2. Surprisingly, values of correlation were quite small, especially for populations with few relatives, although with abundant molecular information figures rise up to 92%. This could be due to the fact that a high amount of couples were nonrelated and therefore their co-ancestry equaled zero. In methods reconstructing genealogies like the present one, estimated co-ancestries are restricted to the biological range (i.e. zero to one) and, thus, any error on nonrelatives comparisons is an overestimation, leading to poor correlations. Probably, this is the reason why studies about co-ancestry estimators do not usually present this parameter as a measure of the accuracy of the different methods (but see Toro *et al*. 2002 for pairwise estimators).

As it would be expected, root mean square errors (RMSE) decreased with increased amount of marker information. In general, RMSE values were small, being comparable or even lower than those reported by Milligan (2003) for different estimators, which ranged from 0.05 to 0.15 for unrelated and full-sibs pairs.

Results for situations where true genealogies comprised more than one generation and different numbers of generations were accounted for by the algorithm are summarized in Table 3 for two particular scenarios (five markers with five alleles each and 20 biallelic markers) and the three types of populations considered.

When the method reconstructed single-generation genealogies, accuracy decreased as the number of generations in the true genealogy increased, at least for highly related populations. For example, with five markers with five alleles and 10 parents, correlation fell from 0.410 to 0.377 and 0.304 when true genealogy had one, two or four generations, respectively. The same behaviour was observed when the parameter for comparison was RMSE (0.074, 0.094 and 0.159 in the above case). When the number of ancestors per generation was high, behaviour of correlations and RMSE's was not so consistent.

The effect of the method being able to account for more than one generation can be seen by comparing accuracy when true genealogies comprised four generations and were estimated by constructing one-, two- or four-generation

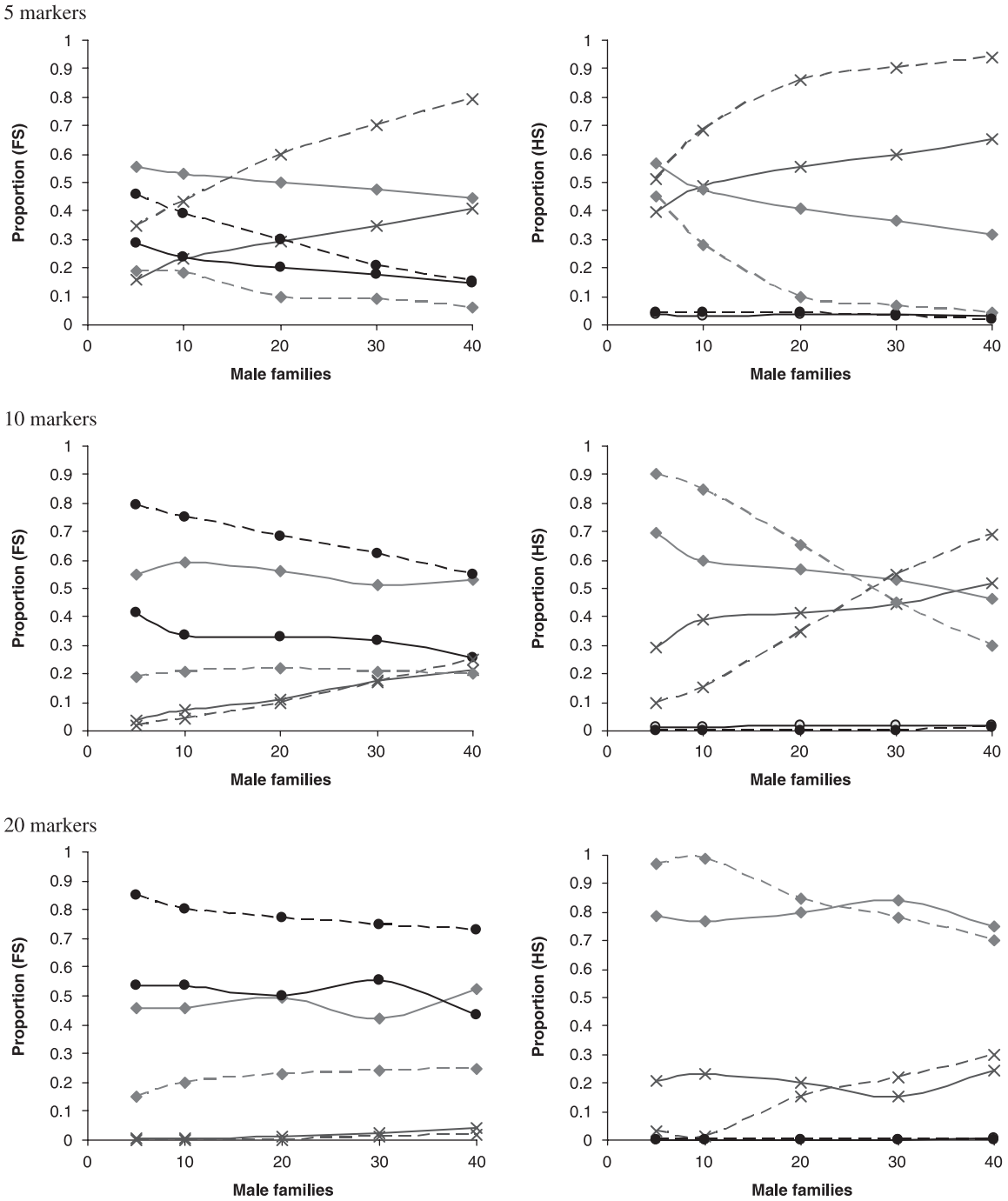

Fig. 1 Proportions of couples assigned to each type of relationship (full-sib, half-sib or unrelated) for different population sizes. Each family comprised two females per male, and each female produces five offspring. All markers have five equally frequent alleles. Left panels: assignment for pairs that are actually full-sibs. Right panels: assignment for pairs that are actually half-sibs. Circles denote proportion of pairs reconstructed as full-sibs, diamonds reconstructed as half-sibs, and crosses reconstructed as nonrelated. Solid lines are results from the present study. Dashed lines are results from Thomas & Hill (2002).

genealogies. With five markers, five alleles per marker and 10 parents, correlation increased from 0.304 to 0.360 and 0.391, respectively, and RSME decreased from 0.159 to 0.141 and 0.116.

It is interesting to point out that correlations increased with increasing number of considered generations (for all simulated scenarios), even when the true genealogy comprised less generations than the estimated one. This effect

Table 1 Number of correct (i.e. actual and estimated co-ancestry coincide) and wrong assignments (standard errors in italic) for unbalanced data sets. Actual relationship is indicated in rows and estimated relationship appears in columns

		5 ♀/5 ♂						20 ♀/20 ♂						50 ♀/50 ♂					
		NR	HS	FS	NR	HS	FS	NR	HS	FS	NR	HS	FS	NR	HS	FS	NR	HS	FS
5 markers		5 alleles			10 alleles			5 alleles			10 alleles			5 alleles			10 alleles		
Actual	NR	2805	358	9	3031	140	1	3869	575	23	4101	360	6	4098	625	27	4349	394	8
		16	*14*	*1*	*10*	*8*	*0*	*16*	*14*	*1*	*15*	*13*	*1*	*19*	*18*	*1*	*11*	*11*	*1*
	HS	966	575	42	877	692	14	292	164	17	269	190	13	121	68	8	108	81	8
		16	*15*	*2*	*28*	*28*	*1*	*5*	*3*	*1*	*5*	*4*	*1*	*2*	*1*	*1*	*3*	*2*	*1*
	FS	54	99	43	36	100	60	3	6	2	2	6	3	1	2	1	0	2	1
		3	*3*	*3*	*3*	*4*	*3*	*0*	*1*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*	*0*
10 markers		5 alleles			10 alleles			5 alleles			10 alleles			5 alleles			10 alleles		
Actual	NR	2981	188	1	3125	45	0	4095	364	6	4327	138	1	4317	431	9	4533	222	2
		19	*19*	*0*	*15*	*15*	*0*	*14*	*12*	*1*	*12*	*10*	*0*	*11*	*11*	*1*	*10*	*9*	*0*
	HS	730	837	15	423	1157	2	256	202	14	194	271	7	101	83	8	74	111	6
		33	*35*	*2*	*56*	*59*	*0*	*5*	*5*	*1*	*5*	*6*	*1*	*3*	*2*	*1*	*2*	*3*	*1*
	FS	20	93	85	4	48	147	2	7	4	1	5	6	0	1	1	0	1	1
		2	*5*	*6*	*1*	*5*	*6*	*0*	*1*	*0*	*0*	*0*	*1*	*0*	*0*	*0*	*0*	*0*	*0*
20 markers		2 alleles						2 alleles						2 alleles					
Actual	NR	2706	433	19				3788	648	31				4003	714	38			
		25	*25*	*3*				*28*	*25*	*2*				*20*	*18*	*2*			
	HS	826	690	76				253	190	28				108	74	11			
		28	*28*	*4*				*6*	*5*	*3*				*3*	*2*	*1*			
	FS	32	99	70				3	7	3				1	1	1			
		3	*4*	*4*				*0*	*1*	*0*				*0*	*0*	*0*			
		5 alleles			10 alleles			5 alleles			10 alleles			5 alleles			10 alleles		
Actual	NR	3128	30	0	3146	12	0	4317	150	1	4447	20	0	4498	255	1	4677	78	0
		12	*8*	*0*	*9*	*4*	*0*	*10*	*10*	*0*	*5*	*2*	*0*	*9*	*9*	*0*	*6*	*5*	*0*
	HS	390	1200	2	246	1346	0	167	296	7	85	384	1	74	112	7	42	148	3
		51	*48*	*0*	*56*	*55*	*0*	*7*	*8*	*1*	*7*	*8*	*0*	*2*	*2*	*1*	*2*	*3*	*0*
	FS	4	38	159	2	11	187	0	4	8	0	2	11	0	1	1	0	1	2
		1	*5*	*5*	*1*	*3*	*4*	*0*	*1*	*1*	*0*	*0*	*1*	*0*	*0*	*0*	*0*	*0*	*0*

Table 2 Correlations between true and estimated co-ancestry matrix and root mean square error (in italic) for unbalanced data sets

	5 ♀/5 ♂			20 ♀/20 ♂			50 ♀/50 ♂		
	2 alleles	5 alleles	10 alleles	2 alleles	5 alleles	10 alleles	2 alleles	5 alleles	10 alleles
5 markers		0.410	0.581		0.225	0.351		0.152	0.264
		0.074	*0.063*		*0.056*	*0.046*		*0.052*	*0.041*
10 markers		0.636	0.835		0.374	0.616		0.257	0.452
		0.059	*0.038*		*0.046*	*0.033*		*0.043*	*0.031*
20 markers	0.468	0.856	0.917	0.269	0.648	0.880	0.166	0.436	0.718
	0.072	*0.036*	*0.024*	*0.058*	*0.032*	*0.018*	*0.056*	*0.033*	*0.020*

could be due to the fact that the input for the algorithm was the molecular co-ancestry matrix. For single-generation genealogies only three types of true relationships exist but for each of them molecular co-ancestry present a variability (e.g. FS pairs may have molecular co-ancestry in a continuous range from zero to one). If we simulated single-generation genealogies, results could be only 0, 0.125 or 0.25. But as the number of simulated generation increased the number

Table 3 Correlations between true and estimated co-ancestry matrix and root mean square error (in italic) for unbalanced data sets and different numbers of generations allowed in the true and/or the estimated genealogy. Generations for estimated genealogies in rows and generations for true genealogies in columns

		5 ♀/5 ♂			20 ♀/20 ♂			50 ♀/50 ♂		
Estimated	True	1 gen	2 gen	4 gen	1 gen	2 gen	4 gen	1 gen	2 gen	4 gen
1 gen	5 markers/5 alleles	0.410	0.377	0.304	0.225	0.221	0.222	0.152	0.165	0.162
		0.074	*0.094*	*0.159*	*0.056*	*0.059*	*0.065*	*0.052*	*0.052*	*0.052*
	20 markers/2 alleles	0.468	0.444	0.411	0.269	0.257	0.244	0.166	0.179	0.159
		0.072	*0.089*	*0.148*	*0.058*	*0.063*	*0.068*	*0.056*	*0.056*	*0.057*
2 gen	5 markers/5 alleles	0.472	0.445	0.360	0.262	0.270	0.267	0.178	0.192	0.192
		0.064	*0.083*	*0.141*	*0.048*	*0.050*	*0.052*	*0.045*	*0.043*	*0.044*
	20 markers/2 alleles	0.484	0.497	0.458	0.292	0.299	0.289	0.186	0.207	0.186
		0.063	*0.076*	*0.137*	*0.049*	*0.049*	*0.050*	*0.044*	*0.044*	*0.043*
4 gen	5 markers/5 alleles	0.501	0.479	0.391	0.275	0.284	0.287	0.187	0.205	0.207
		0.064	*0.066*	*0.116*	*0.057*	*0.055*	*0.050*	*0.055*	*0.054*	*0.050*
	20 markers/2 alleles	0.509	0.525	0.490	0.306	0.313	0.312	0.203	0.216	0.201
		0.065	*0.063*	*0.108*	*0.059*	*0.057*	*0.050*	*0.057*	*0.055*	*0.052*

Table 4 Decrease in the correlation between true and estimated co-ancestry matrix (in percentage respect to the situation with no genotyping errors) for unbalanced data sets when errors in the genotype were generated

		5 ♀/5 ♂			50 ♀/50 ♂		
Genotyping error (%)		2 alleles	5 alleles	10 alleles	2 alleles	5 alleles	10 alleles
1	5 markers		0	3		–1	3
	10 markers		5	10		6	1
	20 markers	3	8	10	1	2	4
5	5 markers		13	12		13	13
	10 markers		17	20		16	13
	20 markers	21	19	23	23	15	10

of classes increased, so the adjustment (correlation) between molecular co-ancestries and estimated ones can be greater and, indirectly, the correlation between true and estimated correlations also improved. For example, with five markers, five alleles per locus and five parents, correlations between estimated and molecular co-ancestries moved from 0.63 to 0.79 and 0.86 when one-, two- or four-generation genealogies were constructed, respectively. The same happened, to some extent, to RMSE, although for extreme situations (i.e. four generations simulated when single generation existed) errors increased.

A measure of confidence on the estimation was calculated, as explained in the Material and methods section, for the simulations with the unbalanced sets. In general, confidence on the estimations was quite high with > 85% of the couples assigned the same co-ancestry in the 1000 best solutions. For the rest of couples, variance of the estimations was small ranging from 0.1 to 12.8×10^{-4}. The same general confidence was observed if we looked at the deviation between the co-ancestry coefficient value for the best solution and the average of the 1000 solutions kept, with mean values < 0.01 (zero values excluded). Notwithstanding, this parameter allowed finding out some doubtful assignations. For example, in one of the simulations the point estimation for the co-ancestry between two particular individuals was 0.25 (i.e. full-sibs) but the average value for the whole set of solutions was 0.19, indicating that in many solutions they were classified as half-sibs.

Genotyping errors

Measures of the robustness of the FT method against genotyping errors are given in Table 4. There, the percentage of reduction in the correlation between real and estimated co-ancestry respect the values obtained with correct data are shown for some combinations of the number of markers and alleles and for two different rates of error (1% and 5%).

As expected, the higher the number of markers the larger the effect on performance, because the probability of

Table 5 Correlations between true and estimated co-ancestry matrix and root mean square error (in italic) for the real data set and different numbers of generations allowed in the estimated genealogy. Column headings stand for the number of markers used in the calculations of molecular co-ancestry

Gen. used	49	18	10
1	0.903	0.915	0.582
	0.156	*0.157*	*0.164*
2	0.945	0.922	0.566
	0.152	*0.126*	*0.164*
3	0.945	0.920	0.582
	0.111	*0.124*	*0.147*
4	0.946	0.922	0.551
	0.119	*0.106*	*0.139*

an individual of carrying an error is higher. The distortion due to the genotyping errors was more important in populations with high levels of co-ancestry because there were less (and smaller) full-sibs families and therefore the probability of detecting and incompatibility for and individual carrying and error is higher. The loss of accuracy for the case of a rate of the 1% was never above the 10%, while for the less probable rate of 5%, decreases were greater (up to 23%). Notwithstanding, more than 70% of the correct estimated relationships in error-free data were still well estimated even when 5% of errors were present (using exclusively half- or full-sibs pairs).

Real data

Results obtained when the estimation method was applied to the real pig data are summarized in Table 5. When genotypes for all of the available markers (49) were used, correlations between real and estimated genealogy were very high (> 0.9), being equal or better than correlations found when using some pairwise estimators (Toro *et al*. 2002). If we used one marker per chromosome (i.e. a total of 18 markers) accuracy recovered the levels obtained with the whole set of markers. However, if we used information on only 10 randomly chosen microsatellites accuracy fell to 0.5.

Estimates of the average co-ancestry ranged from 0.04 when using single-generation genealogies to 0.10 for four-generation genealogies. Pedigree co-ancestry value using the complete genealogy was 0.16. Pairwise estimators tested by Toro *et al*. (2002) on the same data yielded more biased average co-ancestry estimations (besides being negatives). Despite the large number of generations used to calculate the real (genealogical) co-ancestry, RMSEs were not high, being comparable with errors in estimations of single generation's genealogies with other methods (Milligan 2003).

Discussion

Due to the importance of knowing genealogical relationships in different areas of genetics and ecology, many estimators and algorithms have been developed to infer such relationships from molecular marker data when there is a lack of pedigree information. A broad classification of approaches (Butler *et al*. 2004) distinguishes between those reconstructing complete population structure and pairwise methods that do not involve the explicit reconstruction of genealogies. Depending on the assumptions made by each method and the methodology underlying them, most of them suffer from one or several different problems and/or limitations. The most important concerns are (i) results may greatly depend on the knowledge of the true allelic frequencies in the base population; (ii) methods assume that all markers are in Hardy–Weinberg and linkage equilibria and are independent to each other; (iii) pairwise methods can lead to incongruous assignations because they take into account only two individuals at a time; (iv) approaches are usually constructed for particular structured populations (only consider a few relationship classes, e.g. full-sibs or unrelated).

In the present study, we have proposed a simple new approach to estimate relatedness that is free from the limitations above. The method belongs to the first category of estimators and it uses a 'blind search algorithm' (actually *simulated annealing*) to find the feasible genealogy (or population structure) that yield a pedigree-based co-ancestry matrix with the highest correlation with the molecular co-ancestry matrix calculated using the markers. Thus (i and ii) it makes no direct assumptions about allelic frequencies or the distribution of genotypes and haplotypes; (iii) it always provide congruent relationships, as it considers all individuals at a time; (iv) degrees of relatedness can be as complex as desired just increasing the 'depth' (i.e. number of generations) of the proposed genealogies.

Accuracy of the new method (FT) seems to be comparable with that of previous methods. RMSEs of the estimated co-ancestries (see Table 2) ranged from 0.05 to 0.15, values that were equal or lower than those reported for non-related and full-sibs pairs using different pairwise estimators (Milligan 2003). For populations structured in nested half-sibs families, comparisons with one of the available MCMC methods (actually the one described in Thomas & Hill 2000, 2002) give no great differences between methods and, in some situations (i.e. low molecular information and large number of families), FT even outperformed TH method. It must be brought to mind that TH method is specifically designed for nested families and therefore it uses this a priori information. However, in its present form, FT method presumed no familiar structure and has to search in a greater space of solutions. Assignments in unbalanced

populations under FT method were also quite good, detecting the correct co-ancestry for at least 70% of the couples in the most astringent of the situations where five markers with five alleles were genotyped.

A property of the proposed method was the possibility of accounting for more complex structures than full- and half-sibs families. Results presented show that accuracy of estimates improves for genealogies with more than one generation if feasible space also included more than one generation. The 'deeper' the genealogy the larger the space to search in and the concern about the power of the algorithm may arise. However, similar or better results were found when both estimated and true genealogies comprised four generations than using single generation to estimate single-generation genealogies (e.g. 0.287 vs. 0.225 in Table 3 for five markers with five alleles and 40 parents).

Robustness against error in the data is one of the major problems of the estimators involving pedigree reconstruction (see Butler *et al*. 2004). The trouble arises from the fact that all of them consider feasibility/compatibility of full-sibs families and therefore if there is an error in the genotype of one of the individuals it could be erroneously excluded from its family. This effect is greater the highest the number of markers genotyped. Our method also suffers that problem because it checks for compatibility when including and individual into a full-sibs family. However, losses in accuracy seemed to be not as severe as in other methods. Butler *et al*. (2004) tested the robustness of four algorithms against errors concluding that all of them performed quite badly, recovering only from 0 to 70% of the correct assigned individuals when data were error free. Results from the present study (although not directly comparable because they refer to couple instead of individual assignation) show that more than 70% of the correct pair assignments are recovered from the data with error, excluding nonrelated pairs. A possible way to deal with genotyping errors would be to enlarge the feasible space of solutions by relaxing the compatibility rules of full-sibs families. For example, if we have data from 10 markers a particular individual could be included into a family if it is compatible at least for nine of them. This way, we could avoid the misclassification due to the error in the genotype of one locus.

Some authors (see Thomas & Hill 2002) pointed out the fact that MCMC-based methods are conservative in nature and therefore they tend to produce underestimates. The advantage or disadvantage of such a trend depends on the final objective of our study. If using estimates to avoid mating between relatives, overestimates would be preferred, but the opposite would be true if used to estimate breeding values, for example. The method we proposed does not seem to have a clear tendency, but it produces over- or underestimates depending on the level of parentage existing in the population.

No articles in the literature but Toro *et al*. (2002) use the correlation between the true and estimated co-ancestry values as a measure of accuracy. This could be due to the poor values found for populations with low levels of relatedness when estimations are forced to lie in the biological range of [0−1], as explained in the Results section. Notwithstanding, this measure could be important depending on the particular use we make of the relationships estimates. For example, if we are using the co-ancestry matrix to manage the individuals' contributions in a conserved population in order to minimize the loss of genetic diversity, probably it is more important the correct ranking of individuals based on their average co-ancestry that the absolute value of that co-ancestry. Therefore, an estimator yielding large correlations between estimates and true values would work well for that objective, although estimates were biased. A more detailed study should be performed to determine the effect of bias and correlation on the results obtained when co-ancestry estimates are used for different purposes (e.g. minimum co-ancestry mating design, heritability estimation, breeding values estimation, …).

If prior information on the genealogy was available, this could be easily included in the present algorithm. For example, if we knew that a group of individuals share the same mother, we would assign the same virtual mother to all of them from the beginning (i.e. the initial solution) and avoid changing that parent in the random search across the feasible space of solutions (or change it at the same time for that subset of individuals). This scenario is common in plant or forest species that produce fruits carrying several seeds, or in animals that lay eggs in batches (e.g. squids). An extreme situation corresponds to the case of one of the parents known for all individuals in the population, usually the mother. In that case, movements in the *simulated annealing* algorithm would involve exchanging exclusively parents of one sex. Similarly, if an estimation of the number of parents is available, the size of the space of solutions to search in could be reduced increasing the power of the algorithm to detect the optimum configuration.

Information on the structure of the population could also be taken into account by our method. If we knew that only full-sibs families exist, feasible solutions reduces to assigning a number of family to each individual and the algorithm would perform the search by randomly changing individuals from one family to another. In general, prior information could be included by modifying the way initial solution is constructed and the rules that drive the movement from one solution to another.

An extra/parallel use of our method would be to transform/correct co-ancestries estimated through pairwise methods. When real allelic frequencies in the base population are known or well estimated we could use the co-ancestry matrix constructed with the pairwise

estimations as the input of the algorithm, instead of the raw molecular co-ancestry, recovering a solution highly correlated with original coancestries but free from any incongruity, as it would be based on a consistent genealogy. Moreover, no need to define thresholds to decide the values of co-ancestry which separate categories (e.g. pairs are nonrelated if below that value and sibs if above it) because categories would be defined by the pedigree itself.

All simulations have been run assuming independently segregating markers that were under Hardy–Weinberg and linkage equilibria in the base population. The new method makes no assumptions about this matter (in opposition to other estimators) and, probably, would be more robust against deviations from equilibrium situations. However, this assertion deserves a specific study to test for the effect of deviations on different estimators under a wide range of scenarios.

FORTRAN code (or a compiled file) of the algorithm is available from the corresponding author.

Acknowledgements

Authors want to thank L. Silio and the pig-breeding group from the Animal Breeding Department of the INIA for genotypes of pigs from their experimental herd and two anonymous referees for helpful comments on the manuscript. This work was supported by grants BOS2003-03022 from the Ministerio de Ciencia y Tecnología and CPE03-004-C2 from INIA. J. F. was supported by a contract from Programa Ramón y Cajal.

References

Almudevar A (2003) A simulated annealing algorithm for maximum likelihood pedigree reconstruction. *Theoretical Population Biology*, **63**, 63–75.

Almudevar A, Field C (1999) Estimation of single-generation sibling relationships based on DNA markers. *Journal of Agricultural, Biological, and Environmental Statistics*, **4**, 136–165.

Avise JC (1994) *Molecular Markers, Natural History and Evolution*. Chapman & Hall, New York.

Butler K, Field C, Herbinger CM, Smith BR (2004) Accuracy, efficiency and robustness of four algorithms allowing full sibship reconstruction from DNA marker data. *Molecular Ecology*, **13**, 1589–1600.

Caballero A, Toro MA (2000) Interrelations between effective population size and other pedigree tools for the management of conserved populations. *Genetical Research*, **75**, 331–343.

Caballero A, Toro MA (2002) Analysis of genetic diversity for the management of conserved subdivided populations. *Conservation Genetics*, **3**, 289–299.

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

Falconer DS, Mackay Trudy FC (1996) *An Introduction to Quantitative Genetics*, 4th edn. Longman, Harlow, UK.

Frankham R, Ballou JD, Briscoe DA (2002) *Introduction to Conservation Genetics*. Cambridge University Press, Cambridge, UK.

Herbinger CM, Doyle RW, Taggart CT *et al*. (1997) Family relationships and effective population size in a natural cohort of Atlantic cod (*Gadus morhua*) larvae. *Canadian Journal of Fisheries and Aquatic Sciences*, **54**, 11–18.

Kirpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. *Science*, **220**, 671–680.

Lynch M, Walsh B (1998) *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Sunderland, MA.

Malécot G (1948) *Les Mathématiques de l'hérédité*. Masson, Paris.

Milligan BG (2003) Maximum-likelihood estimation of relatedness. *Genetics*, **163**, 1153–1167.

Mousseau TA, Ritland K, Heath DD (1998) A novel method for estimating heritability using molecular markers. *Heredity*, **80**, 218–224.

Ritland K (1996) Estimators for pairwise relatedness and individual inbreeding coefficients. *Genetical Research*, **67**, 175–185.

Smith BR, Herbinger CM, Merry HR (2001) Accurate partition of individuals into full-sib families from genetic data without parental information. *Genetics*, **158**, 1329–1338.

Thomas SC, Hill WG (2000) Estimating quantitative genetic parameters using sibships reconstructed from marker data. *Genetics*, **155**, 1961–1972.

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

Toro M, Barragán C, Óvilo C, Rodrigáñez J, Rodríguez C, Silió L (2002) Estimation of co-ancestry in Iberian pigs using molecular markers. *Conservation Genetics*, **3**, 309–320.

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

The mean research interest of the authors is the design of strategies for the optimisation of conservation and breeding programmes, using jointly classical and new technologies like molecular markers.
