Genet Sel Evol (1997) 29, 5-23

Original article

The value of using probabilities of gene origin to measure genetic variability in a population

D Boichard [1], L Maignel [1,2], É Verrier [1,2]

[1] *Station de génétique quantitative et appliquée, Institut national de la recherche agronomique, 78352 Jouy-en-Josas cedex;*
[2] *Département des sciences animales, Institut national agronomique Paris-Grignon, 16, rue Claude-Bernard, 75231 Paris cedex 05, France*

(Received 28 January 1996; accepted 14 November 1996)

Summary – The increase in inbreeding can be used to derive the realized effective size of a population. However, this method reflects mainly long term effects of selection choices and is very sensitive to incomplete pedigree information. Three parameters derived from the probabilities of gene origin could be a valuable and complementary alternative. Two of these parameters, the effective number of founders and the effective number of remaining founder genomes, are commonly used in wild populations but are less frequently used by animal breeders. The third method, developed in this paper, provides an effective number of ancestors, accounting for the bottlenecks in a pedigree. These parameters are illustrated and compared with simple examples, in a simulated population, and in three large French bovine populations. Their properties, their relationship with the effective population size, and their possible applications are discussed.

probability of gene origin / pedigree analysis / effective number of founders / genetic variability / cattle

Résumé – Intérêt des probabilités d'origine de gène pour mesurer la variabilité génétique d'une population. *L'évolution de la consanguinité est le paramètre classiquement utilisé pour mesurer l'évolution de la variabilité génétique d'une population. Toutefois, elle ne traduit que tardivement les choix de sélection, et elle est très sensible à une connaissance imparfaite des généalogies. Trois paramètres dérivés des probabilités d'origine de gène peuvent constituer une alternative intéressante et complémentaire. Deux de ces paramètres, le nombre de fondateurs efficaces et le nombre restant de génomes fondateurs, sont utilisés couramment dans les populations sauvages mais sont peu connus des sélectionneurs. Une troisième méthode, développée dans cet article, vise à estimer le nombre d'ancêtres efficaces en prenant en compte les goulots d'étranglement dans les généalogies. Ces paramètres sont illustrés avec des exemples simples, une population simulée et trois grandes populations bovines françaises. Leurs propriétés, leur relation avec l'effectif génétique et leurs possibilités d'application sont discutées.*

probabilité d'origine de gènes / analyse de généalogies / nombre de fondateurs efficaces / variabilité génétique / bovin

INTRODUCTION

One way to describe genetic variability and its evolution across generations is through the analysis of pedigree information. The trend in inbreeding is undoubtedly the tool most frequently used to quantify the rate of genetic drift. This method relies on the relationship between the increase in inbreeding and decrease in heterozygozity for a given locus in a closed, unselected and panmictic population of finite size (Wright, 1931). However, in domestic animal populations, some drawbacks may arise with this approach. First of all, in most domestic species, the size of the populations and their breeding strategies have been strongly modified over the last 25–40 years. Therefore, in some situations, these populations are not currently under steady-state conditions and the consequences for inbreeding of these recent changes cannot yet be observed. Second, for a given generation, the value of the average coefficient of inbreeding may reflect not only the cumulated effects of genetic drift but also the effect of the mating system, which is rarely strictly panmictic. Thirdly, and this is usually the main practical limitation, the computation of the individual coefficient of inbreeding is very sensitive to the quality of the available pedigree information. In many situations, some information is missing, even for the most recent generations of ancestors, leading to large biases when estimating the rate of inbreeding. Moreover, domestic populations are more or less strongly selected: in this case, the links between inbreeding and genetic variability become complicated, especially because the pattern is different for neutral and selected loci (see Wray et al, 1990, or Verrier et al, 1991, for a discussion).

Another complementary approach, first proposed in an approximate way by Dickson and Lush (1933), is to analyze the probabilities of gene origin (James, 1972; Vu Tien Khang, 1983). In this method, the genetic contributions of the founders, ie the ancestors with unknown parents, of the current population are measured. Although the definition of a founder is also very dependent on the pedigree information, this method assesses how an original gene pool has been maintained across generations. As proposed by Lacy (1989), these founder contributions could be combined to derive a synthetic criterion, the 'founder equivalents', ie, the number of equally contributing founders that would be expected to produce the same level of genetic diversity as in the population under study. MacCluer et al (1986) and Lacy (1989) also proposed to estimate the 'founder genome equivalent', ie the number of equally contributing founders with no random loss of founder alleles in the offspring, that would be expected to produce the same genetic diversity as in the population under study.

The purpose of this paper is three-fold: (1) to present an overview of these methods, well known to wild germplasm specialists, but less frequently used by animal breeders; (2) to present a third approach based on probabilities of gene origin but accounting for bottlenecks in the pedigree; and (3) to compare these three methods to each other and to the classical inbreeding approach. These approaches will be compared using three different methods: very simple and illustrative examples, a simulated complex pedigree, and an example of three actual French cattle breeds representing very different situations in terms of population size and use of artificial insemination.

CONCEPTS AND METHODS

Probability of gene origin and effective number of founders: the classical approach

A gene randomly sampled at any autosomal locus of a given animal has a 0.5 probability of originating from its sire, and a 0.5 probability of originating from its dam. Similarly, it has a 0.25 probability of originating from any of the four possible grandparents. This simple rule, applied to the complete pedigree of the animal, provides the probability that the gene originates from any of its founders (James, 1972). A founder is defined as an ancestor with unknown parents. Note that when an animal has only one known parent, the unknown parent is considered as a founder. If this rule is applied to a population and the probabilities are cumulated by founders, each founder k is characterized by its expected contribution q_k to the gene pool of the population, ie, the probability that a gene randomly sampled in this population originates from founder k. An algorithm to obtain the vector of probabilities is presented in *Appendix A*. By definition, the f founders contribute to the complete population under study without redundancy and the probabilities of gene origin q_k over all founders sum to one.

The preservation of the genetic diversity from the founders to the present population may be measured by the balance of the founder contributions. As proposed by Lacy (1989) and Rochambeau et al (1989), and by analogy with the effective number of alleles in a population (Crow and Kimura, 1970), this balance may be measured by an effective number of founders f_e or by a 'founder equivalent' (Lacy, 1989), ie, the number of equally contributing founders that would be expected to produce the same genetic diversity as in the population under study

$$f_e = 1 \Big/ \sum_{k=1}^{f} q_k^2 \qquad [1]$$

When each founder has the same expected contribution $(1/f)$, the effective number of founders is equal to the actual number of founders. In any other situation, the effective number of founders is smaller than the actual number of founders. The more balanced the expected contributions of the founders, the higher the effective number of founders.

Estimation of the effective number of ancestors

An important limitation of the previous approach is that it ignores the potential bottlenecks in the pedigree. Let us consider a simple example where the population under study is simply a set of full-sibs born from two unrelated parents. Obviously, the effective number of ancestors is two (the two parents), whereas the effective number of founders computed by equation [1] is four when the grandparents are considered, and is multiplied by two for each additional generation traced. This overestimation is particularly strong in very intensive selection programs, when the germplasm of a limited number of breeding animals is widely spread, for instance by artificial insemination.

To overcome this problem, we propose to find the minimum number of ancestors (founders or not) necessary to explain the complete genetic diversity of the population under study. Ancestors are chosen on the basis of their expected genetic contribution. However, as these ancestors may not be founders, they may be related and their expected contributions q_k could be redundant and may sum to more than one. Consequently, only the marginal contribution (p_k) of an ancestor, ie, the contribution not yet explained by the other ancestors, should be considered. We now present an approximate method to compute the marginal contribution (p_k) of each ancestor and to find the smallest set of ancestors. The ancestors contributing the most to the population are chosen one by one in an iterative procedure. A detailed algorithm is presented in *Appendix B*. The first major ancestor is found on the basis of its raw expected genetic contribution ($p_k = q_k$). At round n, the nth major ancestor is found on the basis of its marginal contribution (p_k), defined as the genetic contribution of ancestor k, not yet explained by the $n-1$ already selected ancestors.

To derive p_k from q_k, redundancies should be eliminated. Two kinds of redundancies may occur. (1) Some of the $n-1$ already selected ancestors may be ancestor of individual k. Therefore p_k is adjusted for the expected genetic contributions a_i of these $n-1$ selected ancestors to individual k (on the basis of the current updated pedigree, see below):

$$p_k = q_k \left(1 - \sum_{i=1}^{n-1} a_i\right)$$

(2) some of the $n-1$ already selected ancestors may descend from individual k. As their contributions are already accounted for, they should not be attributed to individual k. Therefore, after each major ancestor is found, its pedigree information (sire and dam identification) is deleted, so that it becomes a 'pseudo founder'. As mentioned above, the pedigree information is updated at each round. Such a procedure also eliminates collateral redundancies and the marginal contributions over all ancestors sum to one. The number of ancestors with a positive contribution is less than or equal to the total number of founders.

The numerical example presented in table I and figure 1 illustrates these rules. At round 2, after individual 7 has been selected, the marginal contribution of individual 6 is zero because it contributed only through 7, and the pedigree of individual 7 has been deleted. At round 4, after individual 2 has been selected, the marginal contribution of individual 5 is only 0.05 (ie, 0.25 genome of the population under study) because the pedigree of 7 has been deleted and half the remaining contribution of 5 is already explained by 2.

Again, formula [1] could be applied to these marginal contributions (p_k) to determine the effective number of ancestors (f_a)

$$f_a = 1 \Big/ \sum_{k=1}^{f} p_k^2$$

An exact computation of f_a, however, requires the determination of every ancestor with a non-zero contribution, which would be very demanding in large populations.

Table I. Computation of the marginal contribution of the ancestors in the population presented in figure 1 (in bold, the marginal contribution of the selected ancestors).

Ancestors	*Step 1*	*Step 2*	*Step 3*	*Step 4*	*Step 5*	*Step 6*	*Step 7*	*Marginal contribution*
1	0.100	0.100	0.100	**0.100**	–	–	–	0.100
2	0.225	0.150	**0.150**	–	–	–	–	0.150
3	0.125	0.050	0.050	0.050	0.050	0.050	**0.000**	0.000
4	0.100	0.100	0.100	0.100	**0.100**	–	–	0.100
5	0.250	0.100	0.100	0.050	0.050	**0.050**	–	0.050
6	0.150	0.000	0.000	0.000	0.000	0.000	**0.000**	0.000
7	**0.300**	–	–	–	–	–	–	0.300
8	0.300	**0.300**	–	–	–	–	–	0.300
Total								1.000

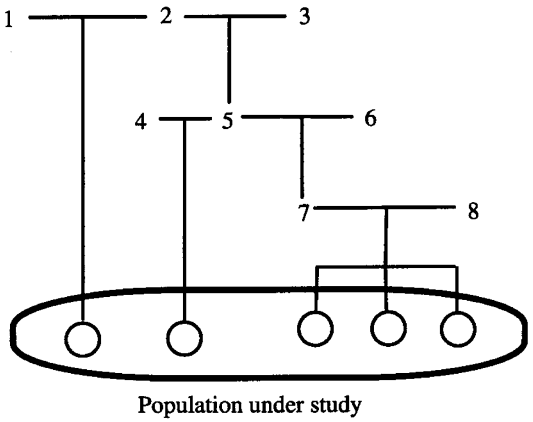

Fig 1. Pedigree used to illustrate the computation of the marginal contributions of the ancestors.

Alternatively, the first n most important contributors could be used to define a lower bound (f_l) and an upper bound (f_u) of the true value of the effective number of ancestors. Let $c = \sum_{i=1}^{n} p_i$ be the cumulated probability of gene origin explained by the first n ancestors, and $1 - c$ be the remaining part due to the other unknown ancestors. The upper bound could be defined by assuming that $1 - c$ is equally distributed over all possible $(f - n)$ remaining founders

$$f_u = \frac{1}{\left[\sum_{i=1}^{n} p_i^2 + \frac{(1-c)^2}{f-n}\right]}$$

Conversely, the lower bound could be defined by assuming that $1-c$ is concentrated over only m founders with the same contribution equal to p_n, and that the contributions of the other ancestors is zero. Consequently, $m = (1-c)/p_n$ and

$$f_l = \frac{1}{\left[\sum_{i=1}^{n} p_i^2 + m p_n^2\right]}$$

As f_l and f_u are functions of n, the computations could be stopped when $f_u - f_l$ is small enough.

This second way of analyzing the probabilities of gene origin presents some drawbacks, however. This method still underestimates the probability of gene loss by drift from the ancestors to the population under study, and, as a result, the effective number of ancestors may be overestimated. Second, the way to compute it provides only an approximation. Because some pedigree information is deleted, two related selected ancestors may be considered as not or less related. Moreover, as pointed out by Thompson (pers comm), when two related ancestors have the same marginal contribution, the final result may depend on the chosen one. However, for the large pedigree files used in this study and presented later on, the estimation of f_a was found to be very robust to changes in the selection order of ancestors with similar contributions p_k.

Estimation of the effective number of founder genes or founder genomes still present in the population under study (Chevalet and Rochambeau, 1986; MacCluer et al, 1986; Lacy, 1989)

A third method is to analyze the probability that a given gene present in the founders, ie, a 'founder gene', is still present in the population under study. This can be estimated from the probabilities of gene origin and by accounting for probabilities of identity situations (Chevalet and Rochambeau, 1986) or probabilities of loss during segregations (Lacy, 1989). However, in a complex pedigree, an analytical derivation is rather complex or not even feasible. MacCluer et al (1986) proposed to use Monte-Carlo simulation to estimate the probability of a founder gene remaining present in the population under study. At a given locus, each founder is characterized by its two genes and $2f$ founder genes are generated. Then the segregation is simulated throughout the complete pedigree and the genotype of each progeny is generated by randomly sampling one allele from each parent. Gene frequencies f_k are determined by gene counting in the population under study. The effective number of founder genes N_a in the population under study is obtained as an effective number of alleles (Crow and Kimura, 1970):

$$N_a = 1 \Big/ \sum_{k=1}^{2f} f_k^2$$

As a founder carries two genes, the effective number of founder genomes (called 'founder genome equivalent' by Lacy, 1989) still present in the population under

study (N_g) is simply half the effective number of founder genes

$$N_g = \frac{N_a}{2} = 1/2 \sum_{k=1}^{2f} f_k^2 \quad [2]$$

N_g seems to be more convenient than N_a because it can be directly compared with the previous parameters (f_e and f_a). This Monte-Carlo procedure is replicated to obtain an accurate estimate of the parameter of interest.

Illustration using a simple example

The simple population presented in figure 2 includes two independent families. Results pertaining to the three methods are presented in table II, for each separate family and for the whole population. The effective number of founders, which only accounts for the variability of the founder expected contributions, provides the largest estimates. In both families, the effective number of founders equals the total number of founders, because all founders contribute equally within each family. This is no longer the case, however, in the whole population, because the founder contributions are not balanced across families. The effective number of ancestors, which accounts for bottlenecks in the pedigree, provides an intermediate estimate, whereas the effective number of founder genomes remaining in the reference population is the smallest estimate, because it also accounts for all additional random losses of genes during the segregations. In family 1, the effective number of founders is higher than the effective number of ancestors, because of the bottleneck in generation 2. The effective number of founder genomes is rather close to the effective number of ancestors, because of the large number of progeny in the last generation, ensuring almost balanced gene frequencies. In contrast, in family 2, the effective number of founders is close to the effective number of ancestors because of the absence of any clear bottleneck in the pedigree, but the effective number of founder genomes is low because of the large probability of gene loss in the last generation. Finally, it could be noted that the estimates are not additive, and the results at the population level are always lower than the sum of the within-family estimates, reflecting unequal family sizes.

COMPARISON OF THESE CRITERIA WITH INBREEDING IN THE CASE OF A COMPLETE OR INCOMPLETE PEDIGREE

Lacy (1989) pointed out there is no clear relationship between the effective size derived from inbreeding trend and the different parameters derived from the probability of gene origin. The goal of this section is simply to compare the robustness of the different estimators proposed in regard to the pedigree completeness level. A simple population was simulated with six or ten separate generations. At each generation, n_m (5 or 25) sires and n_f (25) dams were selected at random among 50 candidates of each sex and mated at random. Before analysis, pedigree information (sire and dam) was deleted with a probability p_m for males and p_f for females. In all situations, pedigree information was complete in the last generation, ie, each

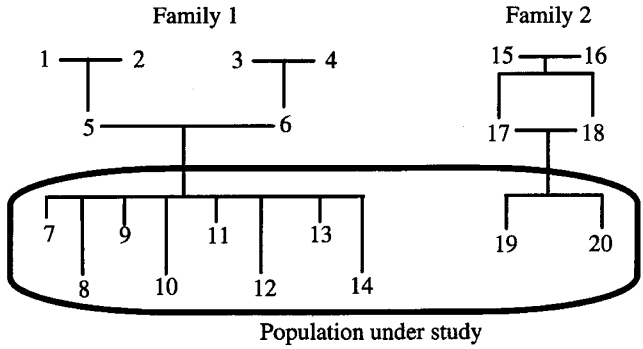

Fig 2. Pedigree used for the comparison of the three methods based on probabilities of gene origin.

Table II. Founder analysis of the pedigree presented in figure 2.

Population	*Total No of founders*	f_e [a]	f_a [b]	N_g [c]
Total	6	5.6	2.94	2.5
Family 1	4	4.0	2.0	1.8
Family 2	2	2.0	2.0	1.1

[a] Effective number of founders. [b] Effective number of ancestors. [c] Effective number of remaining founder genomes.

offspring in this last generation had a known sire and a known dam. Three situations considered were: $p_m = p_f = 0$ (complete pedigree), $p_m = 0$ and $p_f = 0.2$ (the parents of males were assumed to be always known), and ($p_m = p_f = 0.1$). Five hundred replicates were carried out. For founder analysis, the population under study was the whole last generation. For this generation, the effective number of founders (f_e), the effective number of ancestors (f_a), and the effective number of founder genomes (N_g) were computed for each replicate, and averaged over all the replicates. At each generation, the average coefficient of inbreeding was computed. The trend in inbreeding was found to be very unstable from one replicate to another, especially when the pedigree was not complete. In such a situation, the change in inbreeding for a given replicate did not allow us to properly estimate the realized effective size (Ne) of the population. Therefore Ne was only estimated on the basis of results averaged over replicates, using the following procedure. The effective size at a given generation t (Ne_t) was computed according to the classical formula:

$$\frac{1}{Ne_t} = 2\frac{\overline{F}_{t+1} - \overline{F}_t}{1 - \overline{F}_t}$$

where $\overline{F}_t$ is the mean over replicates of the average coefficient of inbreeding at generation t. Next, Ne was computed as the harmonic mean of the observed values

of Ne_t during the last four generations, ie, Ne_2–Ne_5, or Ne_6–Ne_9, when six or ten generations were simulated, respectively.

The results for a population managed over 6 or 10 generations are presented in tables III and IV, respectively. When the pedigree information was complete, the realized effective size was very close to its theoretical value ($4/Ne = 1/n_m + 1/n_f$), as expected. On the other hand, when the pedigree information was incomplete, the computed inbreeding was biased downwards and the realized effective size was overestimated. This phenomenon was particularly clear when considering the long term results. After six generations, the realized effective size with an incomplete pedigree was about twice the effective size with a complete pedigree. After ten generations, it was equal to 3.4–4.2 times the effective size for a complete pedigree and became virtually meaningless. It should be noted that Ne was slightly less overestimated in the case where both the paternal and maternal sides were affected by a lack of information at the same rate than in the case where only the maternal side was affected but at twice as high a rate. In fact, even when n_m equals n_f, a sire–common ancestor–dam pathway is more likely to be cut when the lack of information is more pronounced in one sex.

Table III. Simulation results after six generations.

No sires	*No dams*	*Pr (sire unknown)*	*Pr (dam unknown)*	*Theoretical* Ne[a]	F_6[b]	*Realized* Ne[c]	f_e[d]	f_a[d]	N_g[d]
5	25	0.0	0.0	16.7	0.137	17.2	9.6	8.0	3.2
5	25	0.1	0.1	16.7	0.077	34.9	13.8	10.7	5.0
5	25	0.0	0.2	16.7	0.078	34.4	16.4	11.5	4.9
25	25	0.0	0.0	50.0	0.049	50.4	26.5	22.7	8.2
25	25	0.1	0.1	50.0	0.027	99.1	38.0	29.7	12.4
25	25	0.0	0.2	50.0	0.026	100.7	38.0	29.8	12.4

[a] $4/Ne = 1/No$ sires + $1/No$ dams. [b] Average coefficient of inbreeding at generation 6 (mean of 500 replicates). [c] Effective size derived from the inbreeding rate averaged over replicates (see text). [d] Defined in table II (mean of 500 replicates).

Table IV. Simulation results after ten generations.

No sires	*No dams*	*Pr (sire unknown)*	*Pr (dam unknown)*	*Theoretical* Ne[a]	F_{10}[b]	*Realized* Ne[a]	f_e[a]	f_a[a]	N_g[a]
5	25	0.0	0.0	16.7	0.234	17.1	9.5	7.2	2.1
5	25	0.1	0.1	16.7	0.100	63.7	16.3	10.7	4.3
5	25	0.0	0.2	16.7	0.095	69.6	21.2	11.2	4.4
25	25	0.0	0.0	50.0	0.086	50.8	26.3	21.0	5.3
25	25	0.1	0.1	50.0	0.035	171.8	44.1	29.9	10.6
25	25	0.0	0.2	50.0	0.034	192.5	44.4	30.0	10.7

[a] See tables II and III. [b] Average coefficient of inbreeding at generation 10 (mean of 500 replicates).

Table V. Presentation of the three cattle populations.

Breeds	Abondance	Normande	Limousine
Total No of animals in the pedigree file	106 520	2 338 305	919 561
Total No of animals in the population under study[a]	9 971	301 402	93 591
Average No of ancestors	62.4	159.3	182.3
Complete generation equivalent[b]	3.84	5.02	4.87
Maximum No of generations traced	13	16	18
% known ancestors at generation 2	84.8	90.2	89.3
% known ancestors at generation 4	58.1	77.8	72.8
% known ancestors at generation 6	20.0	44.8	39.8
% known ancestors at generation 8	2.0	11.7	12.9
% known ancestors at generation 10	< 0.1	0.4	1.5

[a] Number N of recorded females born in 1988–91 from known sire and dam.

[b] $\frac{1}{N}\sum_{j=1}^{N}\sum_{i=1}^{n_j}\frac{1}{2^{g_{ij}}}$ with n_j the total number of ancestors of animal j in the population under study and g_{ij} the number of generations between j and its ancestor i.

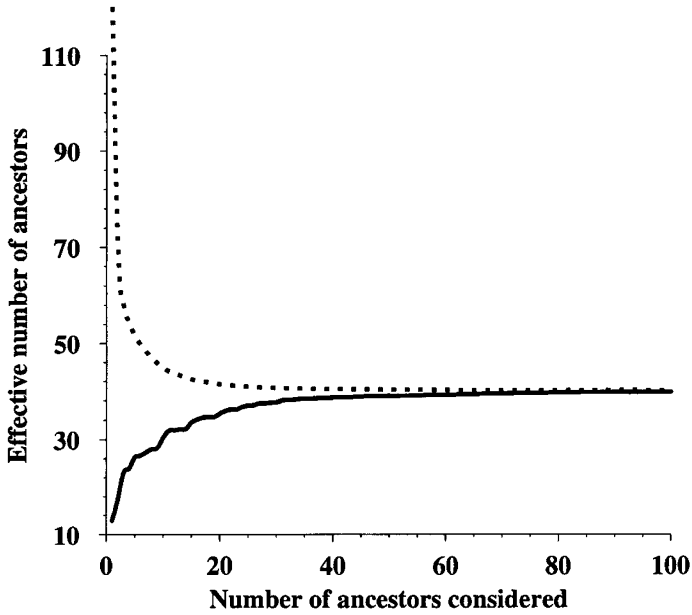

Fig 3. Estimation of the effective number of ancestors (f_a) in the *Normande* breed, according to the number of important ancestors detected. The ancestors are ranked by decreasing contribution. - - -, upper bound; —, lower bound.

The results for the parameters derived from probabilities of gene origin showed a different pattern. First, when the pedigree was complete, the computed values

were, as expected, significantly smaller after ten generations than after six, which was obviously not the case for the effective size. Basically, the three parameters considered (f_e, f_a and N_g) account for the chance of gene loss, which increases with the number of generations. The value of f_e, however, was only slightly affected. The values computed for f_e, f_a and N_g at the tenth generation were equal to around 98, 90 and 64% of the values computed for the sixth generation, respectively. Since f_e refers only to the founders' contributions, it was the least reduced. Conversely, since N_g accounts for all possibilities of founder gene losses, it was the most reduced. Since f_a only accounts for gene losses due to bottlenecks, it was intermediate between the other two parameters. Second, when the pedigree was not complete, these parameters were also affected, but to a smaller extent than the effective size. At the sixth generation, f_e, f_a and N_g were overestimated by 47–72%, 36–45% and 57%, respectively. At the tenth generation, the amount of overestimation was of the same magnitude, or a bit smaller: 45, 32 and 54%, respectively. Although they were consistently biased, these parameters, and particularly f_a, appeared to be more robust to partial lack of pedigree information than the realized effective size. Interestingly, with an incomplete pedigree, f_e was larger at generation 10 than at generation 6, due to the larger number of false founders.

APPLICATION TO THREE LARGE CATTLE PEDIGREE FILES

Three populations were considered, representing three different but typical situations. The *Abondance* breed is a red-and-white dairy breed originating from and located in the northern French Alps. It is of limited population size, with about 3 000 new heifers milk recorded each year and 106 520 animals in the whole pedigree file. The *Normande* breed is a dairy population located in the northwestern half of France. It has quite a large population size, with about 80 000 new heifers milk recorded each year, and 2 338 305 animals in the pedigree file. The *Limousine* breed is a beef population located in the western part of the Massif Central mountains. It is of intermediate population size, with about 25 000 new registered heifers each year and 919 561 animals in the pedigree file. Both dairy breeds are characterized by the predominant use of a limited number of bulls widely spread by artificial insemination. In contrast, the beef breed uses mainly natural matings, with only 15% artificial insemination. More detailed results, including all the main French dairy breeds, will be presented elsewhere.

The pedigree information was better in *Limousine* and *Normande* than in *Abondance* breed. It was best in the *Normande* population in the first seven generations and in the *Limousine* in the older generations (table V). However, the pedigree should be considered as incomplete because only 78 and 45% of ancestors were known at generations 4 and 6, respectively, in the best situation, ie, the *Normande* one. The population under study was defined by all females born between 1988 and 1991 from known sires and dams. Consequently, it included an almost complete generation. The parameters f_e, f_a and N_g were computed as described previously. For the computation of f_a, the process was stopped in *Abondance* and *Normande* when the 100 most important ancestors were detected. This corresponded to very little difference between the lower and upper bounds of f_a, as illustrated in figure 3. In *Limousine*, 500 ancestors were required to reach

a sufficient level of accuracy. Individual coefficients of inbreeding were computed according to the method proposed by VanRaden (1992). Although this method is less efficient than that of Meuwissen and Luo (1992), it has been preferred here because it makes it possible to assume that the founders are not independent and, therefore, to some extent can accomodate incomplete pedigree information. VanRaden's method is derived from the classical tabular method applied to each individual and all its ancestors. Each unknown ancestor is put into a group according to its birth year. The first rows and columns of the table are dedicated to the groups. The group by group subtable includes the average relationship coefficients within and between groups of founders. It is initialized by values computed iteratively. At the first run, zeros are used as starting values. At the next rounds, the following rules were used. Within a given group, the average relationship coefficient among founders born in a given year was assumed to be twice the average inbreeding coefficient of the animals with known parents and grandparents and born 5 years (ie, close to one generation) later. The relationship coefficient between founders from different groups was assumed to be equal to the relationship coefficient within the most recent group. In practice, convergence was reached after three rounds. In comparison with assuming no relationship between founders, this procedure led to a 20% higher inbreeding level in the population of *Normande* females born in 1988–91. The effective size of the populations (Ne) was estimated from the average increase in inbreeding during the last generation for the animals with known parents and grandparents.

The results are presented in table VI. Inbreeding presented a very different pattern from one breed to another. A strong increase of more than 1% per generation was observed in the *Normande* breed, a moderate increase in the *Abondance* breed, and a decrease in the *Limousine*. Accordingly, the effective size was the smallest in the *Normande* breed (47), while it was not estimable in the *Limousine*. These results illustrated the difficulty of using inbreeding to quantify the genetic drift within a population when the pedigree information is incomplete and when only a few generations of animals are available in the pedigree file.

Table VI. Results of the pedigree analysis in the cattle populations.

Breeds	*Abondance*	*Normande*	*Limousine*
Inbreeding increase[a]	0.47	1.07	−0.05
Effective size (Ne)[b]	106	47	–
Total number of founders (f)	6 109	138 291	26 656
Effective No of founders (f_e)	69	132	790
Effective No of ancestors (f_a)	25	40	360
Effective No of founder genomes (N_g)	17	22	206

[a] In %, during the last generation. [b] Derived from inbreeding rate.

In contrast, the probability of gene origin provided results that were more convincing and easier to interpret. The effective number of founders (790) was highest in *Limousine*, because of the predominance of natural mating, and lowest in *Abondance*, because of artificial insemination and its small population size. However,

the very limited effective number of founders (132) of the *Normande* breed shows that the breeding system and the effective number of sires were more determinant than the number of females. Whereas the f_e/f_a ratio was only 2 in the *Limousine*, it reached 3 in both dairy breeds, illustrating the narrower bottlenecks in populations where artificial insemination is widely used. The very small effective number of ancestors in *Abondance* and *Normande*, 25 and 40, respectively, could be illustrated by the number of ancestors required to explain 50% of the genes, which was found to be only 8 and 17, respectively. Finally, the effective number of founder genomes remaining in the reference group was even lower, 17, 22 and 206 in *Abondance*, *Normande*, and *Limousine* populations, respectively. The lowest N_g/f_e ratio was in the *Normande* breed, showing that the genetic drift was greater in this population, probably because the major ancestors were older than in the other breeds.

DISCUSSION AND CONCLUSIONS

Properties of the different parameters

Three parameters based on the probabilities of gene origin are introduced, in addition to the usual effective size based on inbreeding trend. The effective number of founders (f_e) measures how the balance in founder expected contributions is maintained across generations. It accounts for selection rate (ie, the probability of being a parent or not) and for the variation in family size, but it neglects the probability of gene loss from parent to progeny. The effective number of ancestors (f_a) accounts for bottlenecks in the pedigree, which is the major cause of gene loss in some populations, as in dairy cattle. Consequently f_a is always less than or equal to f_e. Finally, the effective number of founder genomes (N_g) measures how many founder genes have been maintained in the population for a given locus, and how balanced their frequencies are. It accounts for all causes of gene loss during segregations and, consequently, provides a smaller number than f_a and f_e.

Although the parameters presented here are related to the effective size, they should not be directly compared to it. One reason lies in the difference in trends over time. The effective size (Ne) is a function of the relative increase in inbreeding or the variance of gene frequency from one generation to another. In a given population with a constant structure, Ne is expected to remain the same across generations. In contrast, f_e, f_a and N_g are expected to decrease over time, particularly N_g which fully accounts for genetic drift, as shown by the simulation results presented here. This phenomenon may also be illustrated by the comparison of two groups of animals within the three cattle breeds analyzed, the females born in 1984–1987 or in 1988–1991 (table VII). Since the time interval between both groups is close to one bovine generation, the relative decrease observed for the three parameters (-10.5 to -21.1%, except -3.6 for f_e in *Normande*) represents a dramatic change in genetic variability. It should be kept in mind, however, that starting from a hypothetical base population, the reduction in f_e, f_a or N_g is rapid by nature, because most gene losses occur very early in the first generations. This phenomenon clearly appears when comparing the values computed for the simulated populations with complete pedigree (tables III and IV) to the total number of founders considered, ie, 30 and 50, respectively. This early loss of genes is a well established result either

analytically (Engels, 1980) or by simulation (Verrier et al, 1994). For a given locus, the number of alleles in a base population is generally much lower than the total number of founder genes, even for very polymorphic loci. As a consequence, the allelic diversity, measured by the effective number of alleles (Crow and Kimura, 1970) for example, is expected to decrease due to drift at a lower rate than the parameters considered here.

Table VII. Evolution over time of the parameters derived from founder analysis in the cattle populations.

Breeds	Abondance		Normande		Limousine	
	(*a*)	(*b*)	(*a*)	(*b*)	(*a*)	(*b*)
Effective No of founders	79	−12.7	137	−3.6	937	−15.7
Effective No of ancestors	29	−13.8	46	−13.0	420	−14.3
Effective No of founder genomes	19	−10.5	26	−15.4	261	−21.1

(a) Value for animals born from 1984 to 1987. (b) Relative difference between values for animals born from 1988 to 1991 (see table VI) and (a), expressed in percentage of (a).

Effective size and parameters derived from probabilities of gene origin, however, are related because they more or less account for the same basic phenomena, ie, unbalanced contributions of parents to the next generation and loss of genes from a given parent to its progeny. Clearly, the smaller Ne, the higher the decrease of N_g over time. This may be shown in a simple way. At a given generation, according to equation [2], the effective number of genomes N_g, is half the effective number of founder genes N_a. Let us define H as the expected rate of heterozygotes in a population under random mating at a locus with N_a alleles and balanced frequencies $(1/N_a)$. Therefore

$$H = 1 - 1/N_a = 1 - 1/2N_g \quad [3]$$

Asymptotically, the rate of decay of H (ΔH) from generation t to $t+1$ depends on the effective population size Ne, according to the following classical formula

$$\Delta H = \frac{H_{t+1} - H_t}{H_t} = \frac{-1}{2Ne} \quad [4]$$

Therefore, by combining equations [3] and [4], one obtains

$$\Delta H = \frac{N_{g_{t+1}} - N_{g_t}}{N_{g_{t+1}}(2N_{g_t} - 1)} = \frac{-1}{2N_e} \quad [5]$$

which could provide an estimation of Ne derived from the evolution of N_g.

Similarly, the smaller Ne, the smaller the ratios f_e/f or f_a/f computed at a given generation. In a more general way, it has been shown (James, 1962), in the case of panmictic and unselected populations, that the effective size based on the change in gene frequencies may be derived from a probability of gene origin approach. In the same way, probabilities of identity by descent and effective sizes may be derived from coalescence times (see, for example, Tavaré, 1984). Obviously,

the parameters presented here are related to coalescence times. For example, a bottleneck in pedigree between the founders and the population under study leads to a reduction in both the average coalescence time and the effective number of ancestors. However, more algebra is required to assess the link between parameters presented here and coalescence times.

When studying real populations, an important property is the sensitivity to incomplete pedigree information. In large domestic animals, the pedigree information is limited, incomplete, and variable across animals. The simulation study shows that the inbreeding trend is well estimated only when the pedigree information is complete. Even with a rather small proportion of unknown pedigrees (10%), inbreeding is strongly underestimated. Parameters derived from the probability of gene origin are also affected, but to a smaller extent. In fact, the robustness is highest for the effective number of ancestors (f_a), because it relies on shorter relationship pathways than the other parameters. In contrast, inbreeding estimation relies on the longest relationship pathways, which are more likely to be affected by a lack of information. For the same reason, robustness also increased for all parameters when the number of generations decrease. Although N_g appeared to be less affected by incomplete pedigree than inbreeding, an indirect prediction of Ne from N_g with equation [5] was not found to be more robust than the classical prediction through the inbreeding trend.

All these parameters are easy to compute. Several efficient algorithms have been recently proposed to compute inbreeding (Meuwissen and Luo, 1992; VanRaden, 1992). As shown in *Appendix A*, the computation of f_e is straightforward. Estimation of N_g only requires a good random number generator. The iterative procedure to obtain f_a may be computationally demanding in large populations without strong bottlenecks, ie, when a large number of ancestors should be detected. However, this parameter is interesting especially when strong bottlenecks do exist in the pedigree structure. In practice, none of the analyses of the cattle populations required more than 10 min of CPU time on a IBM 590 Risc6000 workstation.

Practical use of these parameters

The effective size is a powerful tool for predicting the change in genetic variability over a long time period, when the inbreeding increase fully reflects the number and the choice of breeding animals in the previous generations. In contrast, parameters derived from probability of gene origin are very useful for describing a population structure after a small number of generations. They can characterize a breeding policy or detect recent significant changes in the breeding strategy, before their consequences appear in terms of inbreeding increase. From that point of view, they are very well suited to some large domestic animal populations, which have a variable and limited number of generations traced and which have undergone drastic changes in their breeding policy in the last two decades.

The present paper shows how to use parameters derived from probabilities of gene origin in a retrospective way to analyze the genetic structure of domestic populations. Such an analysis, in addition to the more classical approach based on inbreeding, provides a good view of the basis upon which selection is applied. Some recent studies have been realized in that aspect, eg, with dairy sheep (Barillet

et al, 1989), or in race and riding horses (Moureaux et al, 1996). This approach is particularly useful when the main breeding objective is the maintenance of a given gene pool rather than genetic gain, a situation which occurs in rare breed conservation programmes. When a population has been split into groups for its management, the analysis of gene origins in reference to the foundation groups is definitely the method of choice in order to appreciate the genetic efficiency of the conservation programme (see, for instance, Rochambeau and Chevalet, 1989, Giraudeau et al, 1991 and Djellali et al, 1994). The gene origin approach may also be used in selection experiments analysis (eg, James and McBride, 1958; Rochambeau et al, 1989). In a similar way, when analyzing the consequences of selection in a small population via simulation, the gene origins approach provides results which satisfactorily complete the analysis of the trends of the average coefficient of inbreeding or the genetic variance of the selected trait (eg, Verrier et al, 1994).

When looking at real populations, it is generally useful to predict the evolution of genetic variability. Especially in selected populations, such a prediction is necessary to predict selection response. The effective size allows us to predict the reduction in genetic variance in the next generations, assuming that Ne is well estimated from the past. On the other hand, parameters derived from probabilities of gene origin appear to be more descriptive than predictive. Indirectly, they can be used to derive Ne (see above). Another possible way would be to use the approach of James (1971) by replacing the number of founders by the effective number of founders (or ancestors, or genomes) computed in the population under study. Further investigation is needed in this field.

Finally, these parameters could be used as a selection criterion when managing populations under conservation. Alderson (1991) proposed to compute a vector of gene origin probabilities for each newborn in reference to the founders and its own effective number of founders (f_e), and then to select animals with the highest f_e values. Other simple rules have been previously proposed for the management of captive populations of wild species (eg, Templeton and Read, 1983; Foose, 1983). Obviously, the higher the quality of pedigree information, the more efficient these methods will be for managing the genetic variability within a population.

ACKNOWLEDGMENTS

The authors are grateful to D Laloë who provided the *Limousine* data set, to the anonymous referees for their valuable comments, and to E Thompson for the English revision.

REFERENCES

Alderson GLH (1991) A system to maximize the maintenance of genetic variability in small populations. In: *Genetic Conservation of Domestic Livestock* (L Alderson, I Bodo, eds), CAB International, Wallingford, 18–29

Barillet F, Roussely M, Vu Tien Khang J, Poivey JP, Chevalet C, Elsen JM, Rochambeau H de (1989) Variabilité génétique dans le noyau de sélection des ovins laitiers de race Lacaune. In: *La gestion des ressources génétiques des espèces animales domestiques* (M Molénat, É Verrier, eds), Lavoisier, Paris, 71–80

Chevalet C, Rochambeau H de (1986) Variabilité génétique et contrôle des souches consanguines. *Sci Tech Anim Lab* 11, 251–257

Crow JF, Kimura M (1970) *An Introduction to Population Genetics Theory.* Harper & Row, New York

Dickson WF, Lush JL (1933) Inbreeding and the genetic history of the Rambouillet sheep in America. *J Heredity* 24, 19–33

Djellali A, Vu Tien Khang J, Rochambeau H de, Verrier E (1994) Bilan génétique des programmes de conservation des races ovines Solognote et Mérinos Précoce. *Genet Sel Evol* 26, s255–s265

Engels WR (1980) Loss of selectively neutral alleles in small populations and regular mating systems. *Theor Pop Biol* 17, 345–364

Foose TJ (1983) The relevance of captive populations to the conservation of biotic diversity. In: *Genetics and Conservation* (CM Schonewald-Cox, SM Chambers, B Mac Bryde, WL Thomas, eds), Benjamin/Cummings, London, 374–401

Giraudeau L, Verrier É, Rochambeau H de, Ménissier F, Laloe D, Casane D, Poupinot JP, Bougler J, Vu Tien Khang J (1991) Survey and management of the genetic variability of a cattle breed of small population size: a successful program in the Parthenaise breed. *42th Annual Meeting of the European Association for Animal Production, Berlin, Germany, September 9–12 1991*

James JW (1962) The spread of genes in random mating control populations. *Genet Res* 3, 1–10

James JW (1971) The founder effect and response to artificial selection. *Genet Res* 16, 241–250

James JW (1972) Computation of genetic contributions from pedigrees. *Theor Appl Genet* 42, 272–273

James JW, McBride G (1958) The spread of genes by natural and artificial selection in a closed poultry flock. *J Genet* 56, 55–62

Lacy RC (1989) Analysis of founder representation in pedigrees: founder equivalents and founder genome equivalents. *Zoo Biol* 8, 111–123

MacCluer JW, Van de Berg JL, Read B, Ryder OA (1986) Pedigree analysis by computer simulation. *Zoo Biol* 5, 147–160

Meuwissen THE, Luo Z (1992) Computing inbreeding coefficients in large populations. *Genet Sel Evol* 24, 305–313

Moureaux S, Verrier E, Ricard A, Mériaux JC (1996) Genetic variability within French race and riding horse breeds from genealogical data and blood marker polymorphism. *Genet Sel Evol* 28, 83–102

Rochambeau H de, Chevalet C (1989) Aspects théoriques de la gestion des ressources génétiques: cas des populations en conservation. In: *La gestion des ressources génétiques des espèces animales domestiques* (M Molénat, É Verrier, eds), Lavoisier, Paris, 171–180

Rochambeau H de, La Fuente LF de, Rouvier R, Ouhayoun J (1989) Sélection sur la vitesse de croissance post-sevrage chez le lapin. *Genet Sel Evol* 21, 527–546

Tavaré S (1984) Line of descent and genealogical processes and their implication in population genetics. *Theor Pop Biol* 26, 119–164

Templeton AR, Read B (1983) The elimination of inbreeding depression in a captive herd of Soeke's Gazelle. In: *Genetics and Conservation* (CM Schonewald-Cox, SM Chambers, B Mac Bryde, WL Thomas, eds), Benjamin/Cummings, London, 241–261.

VanRaden PM (1992) Accounting for inbreeding and crossbreeding in genetic evaluation for large populations. *J Dairy Sci* 75, 3136–3144

Verrier É, Colleau JJ, Foulley JL (1991) Methods for predicting genetic response to selection in small populations under additive genetic models: a review. *Livest Prod Sci* 29, 93–114

Verrier É, Colleau JJ, Foulley JL (1994) Genetic variability when selecting on the animal model BLUP. In: *5th World Congress on Genetics Applied to Livestock Production, Guelph, August 7–12 1994*, 19, 139–142

Vu Tien Khang J (1983) Méthodes d'analyse des données démographiques et généalogiques dans les populations d'animaux domestiques. *Genet Sel Evol* 15, 263–298

Wray NR, Woolliams JA, Thompson R (1990) Methods for predicting rates of inbreeding in selected populations. *Theor Appl Genet* 80, 513–521

Wright S (1931) Evolution in Mendelian populations. *Genetics* 16, 97–159

APPENDIX A

A simple algorithm to compute the probabilities of gene origin ($\mathbf{q}$):

(1) define the population under study, ie, the group of N animals carrying the gene pool of interest;

(2) initialize a vector $\mathbf{q}$ with 1 for animals in the population under study, with 0 otherwise;

(3) process the pedigree file from the youngest animal to the oldest animal:

$$\text{if sire } (i) \text{ is known then } q(\text{sire}(i)) = q(\text{sire}(i)) + 0.5^{*}q(i),$$
$$\text{if dam}(i) \text{ is known then } q(\text{dam}(i)) = q(\text{dam}(i)) + 0.5^{*}q(i);$$

(4) if an animal is a 'half founder' (ie, with one known parent and one unknown parent), multiply its contribution by 0.5. This is equivalent to considering the unknown parent as a founder. Divide the vector $\mathbf{q}$ by N, so that founder contributions sum to 1.

APPENDIX B

Algorithm for determining the most important ancestors of a population and their marginal contributions:

(1) define the population under study, ie, the group of N animals carrying the gene pool of interest;

(2) we assume that the first $k-1$ most important ancestors are already found. Note the first one is chosen according to its raw contribution computed as in *Appendix A*.

(3) delete the pedigree information (sire and dam information) for the $k-1$ ancestors already found;

(4) initialize a vector $\mathbf{q}$ with 1 for animals in the population under study, with 0 otherwise, and another vector $\mathbf{a}$ with 1 for the $k-1$ ancestors already selected, and with 0 otherwise;

(5) process the pedigree file from the youngest animal to the oldest animal:

$$\text{if sire}(i) \text{ is known then } q(\text{sire}(i)) = q(\text{sire}(i)) + 0.5^{*}q(i),$$
$$\text{if dam}(i) \text{ is known then } q(\text{dam}(i)) = q(\text{dam}(i)) + 0.5^{*}q(i);$$

(6) process the pedigree file from the oldest animal to the youngest animal:

$$\text{if sire}(i) \text{ is known then } a(i) = a(i) + 0.5^{*}a(\text{sire}(i)),$$
$$\text{if dam}(i) \text{ is known then } a(i) = a(i) + 0.5^{*}a(\text{dam}(i));$$

(7) compute the marginal contribution $p(i)$ of each animal i, defined by the proportion of genes it contributes that are not yet explained by its already selected ancestors. This is done by subtracting the contributions of the ancestors already selected from the probabilities of gene origin

$$p(i) = q(i)^{*}(1 - a(i))$$

(8) select the kth ancestor with the highest p value. Divide this value by N, so that contributions over all ancestors sum to 1;
(9) go to 3 for the next ancestor.