Methodological approaches to estimating effective population size

Authors – Luboš Vostrý, Barbora Hofmanová, Hana Vostrá-Vydrová, Karolína Dvořáková Machová

Study material

The conservation of endangered species is one of the most important goals for the current biological sciences sector, especially in maintaining the natural ecosystem. In the case of domesticated animals, a conservation program is usually initiated in breeds that present unique genetic and phenotypic characteristics. Traditional breeding programs are mainly based on selection to improve economically important traits. The consequence of this selection is to reduce the genetic variability of a given population. Thus, conservation programs must target breeds or species that have conserved significant genetic variability (FAO, 2000). The bottleneck effect is among the main factors influencing the loss of genetic diversity (Vicente et al., 2012). The bottleneck effect occurs when there is a rapid reduction in the number of individuals participating in reproduction, which leads to a significant reduction in population size. This population may be able to recover its population size during the following generations, but due to genetic drift, the genetic diversity of this population may change substantially, i.e., allele frequencies may change (Relichová, 2009).

The problem of genetic diversity has become much discussed in the last few decades due to the increasing industrialization of agriculture and the consequent genetic uniformity of crops and livestock. However, the diversity of livestock breeds is essential for future adaptation to different diseases, production systems, and changing climatic and market conditions (FAO 2000). Therefore, conservation programs focus mainly on endangered local breeds (Notter 1999). These tend to be less well-mapped but generally show far greater genetic diversity than international breeds with unique economically attractive traits (FAO 2000).

We describe biodiversity as the extent of all the diversity found in nature (Mészáros 2018). It can be analysed based on different characteristics: phenotypic (morphological), cytological, biochemical, and molecular (Saravanan et al., 2022). However, genetic diversity refers to the extent of variability (polymorphism) directly in the DNA sequence (Ellegren & Galtier 2016).

From a molecular perspective, there are three main types of variability, namely single nucleotide polymorphisms (SNPs), deletions or insertions of different lengths, and variations in the number and length of repetitive sequences (VNTR) (Vignal et al. 2002). New alleles appear with each successive generation of spontaneous mutations due to replication errors or the action of mutagens, and thus, from a theoretical point of view, genetic diversity can be described as the result of a balance between the loss and gain of these alleles (Ellegren & Galtier 2016). The mutation rate decreases within the different level units of the nuclear genome from single genes to chromosomal-level variation (Hodgkinson & Eyre-Walker 2011), which differs significantly between autosomes and gonosomes (Ellegren & Galtier 2016). Significant differences in mutation rates also exist between nuclear and mitochondrial DNA, between gametes and somatic cell lines, and between different species (Lynch 2010).

Over the last century, cropland productivity has increased by 2% per year due to rapid technological and genetic progress. Consequently, the price of grain and other crops has decreased, making it cheaper for many farmers to feed on grain rather than rely wholly on grazing, dramatically changing entire farming systems. Newly bred, high-yielding animals and advanced breeding technologies have pushed traditional breeding systems to the sidelines because of their higher risk (Mendelsohn 2003). The inappropriate use of the BLUP model of breeding values has exacerbated the situation, leading to rapid loss of genetic diversity. While it maximizes the response to selection, it can lead to the selection of closely related individuals (van Wyk et al. 2009). The selected related individuals then further increase the level of inbreeding in the future (van Wyk et al. 2009). This practice could be better because the success of future production improvement may depend on genetic diversity, which appears neutral under current selection criteria and may disappear completely due to a mere failure to document utility (Tapio et al. 2006a).

With the introduction of genomic selection, it was logically assumed that this consequence would disappear because genomic selection based on the whole genome should emphasize differences between siblings and thus reduce the chance of their simultaneous selection based on a more accurate estimate of the Mendelian heritability component and thus breeding values (Daetwyler et al., 2007). However, as many studies in cattle show, the opposite is often true (e.g.:Forutan et al., 2018; Scott et al., 2021). As a result of genomic selection, generation interval and mendelian sampling variability are reduced, which again leads to increased inbreeding, reduced effective population size, and random fixation or loss of alleles, whether due to genetic drift or selection (Makanjuola et al., 2020b). Therefore, an alternative approach to genomic selection based on the principle of contributional selection has been proposed to keep inbreeding under control - optimal contributional genomic selection (Sonesson et al., 2012; Woolliams et al., 2015). The main principle is to maximize genetic response under given levels of inbreeding based on the relatedness of selection candidates and considering their genetic contribution (Dagnachew & Meuwissen, 2016).

Effective population size

Wright (1968) defined effective population size (*Ne*) as the size of an idealized population that can provide an equal increase in the coefficient of inbreeding or the rate of change in the variability of allele frequencies in an observed population. This concept is a basic parameter used as a criteria for determining endangered status, not only in livestock (FAO, 2000). As Falconer and Mackay (1998) state, *Ne* is considered a basic parameter because of the relationship between *Ne* and the increase in inbreeding, fitness level and between the loss of genetic variability due to random genetic drift.

Effective population size and idealised population (Caballero, 1994)

In an infinitely large population and in the absence of mutation, migration, and selection, the frequency of alleles and genotypes stays constant across generations. However, in a finite population, allele frequencies fluctuate randomly from generation to generation due to limited gene selection. This phenomenon is called dispersal or *genetic drift*. Due to this genetic drift, alleles become fixed in the population. Genetic drift can be evaluated in unstructured populations by a simple parameter such as the *effective population size (Ne)*, which can be estimated even under field conditions. As already mentioned, the simplest conditions under which the dispersal process can be studied is an *idealized population, which* includes an infinite, randomly mating population that can be divided into infinitely many subpopulations. Each subpopulation includes a constant number of mating individuals (*N*) per generation. In each subpopulation, parents produce infinitely many male and female gametes, of which only *2N* gametes will fuse to produce *N* zygotes of the subsequent generation. In an idealized population, all individuals survive from zygote to adult, and each individual has an equal probability of producing offspring. In this idealized population, no systematic allele frequency changes, overlapping generations, and only autosomal loci are considered.

In this idealized population, dispersal processes such as gamete selection or inbreeding can be observed, as both phenomena increase the variability of allele frequency between subpopulations.

Under these idealized conditions, gamete selection exhibits a binary distribution, and the variance of the change in genetic variability can be expressed as:

$$\sigma^2_{\Delta q} = \frac{q(1-q)}{2N}, \tag{1}$$

where *q* is the frequency of alleles in the infinite population, the coefficient of inbreeding in generation *t* can then be derived from a relationship where the first part represents the copies of genes by descent of individuals in generation *t-1* and the second part represents the copies of genes of individuals in the previous generation. The increase in the inbreeding coefficient can then be expressed as:

$$\Delta F = \frac{1}{2N}, \tag{2}$$

where

$$\Delta F = \frac{F_t - F_{t-1}}{1 - F_t}. \tag{3}$$

As a result of inbreeding, heterozygosity decreases from generation to generation according to the relationship

$$\lambda = \frac{H_t}{H_{t-1}} = 1 - \Delta F, \tag{4}$$

or relative to the base population

$$\frac{H_t}{H_o} = 1 - F^t = (1 - \Delta F)^t. \tag{5}$$

The relationship between allele variability across subpopulations and the inbreeding coefficient can then be expressed as

$$\sigma^2_{q,t-1} = q(1-q)\left[1 - \left(1 - \frac{1}{2N}\right)^t\right] = q(1-q)F_t, \tag{6}$$

where going back one generation (*t-1*) is because genetic drift starts 1 generation before inbreeding occurs if self mating is considered. If self mating is not considered (in most mammals), genetic drift occurs 2 generations earlier. Thus, when *N* individuals are randomly selected for inbreeding from an infinitely large population, inbreeding is not yet present, but genetic drift has already occurred.

Based on the findings above, the effective population size is defined as the size of the idealized population that provides the increase in variability in allele frequency change or increase in inbreeding that is obtained in the population of interest, i.e.:

$$N_e = \frac{q(1-q)}{2\sigma^2_{\Delta q}}, \text{ nebo } N_e = \frac{1}{2\Delta F}. \tag{7}$$

Thus, *Ne* evaluates the rate of genetic drift and inbreeding variation in a population (Caballero, 1994).

Difference in the number of males and females (Caballero, 1994)

Let us assume differences between the number of males (*Nn*) and females (*Nm*) that are constant over generations. Half of the genes in any generation (e.g., *t-1*) come from the fathers, and the other half come from the mothers. The probability that two alleles in generation *t-1* that merge in population *t* into an offspring (zygote) come from a single individual in generation *t-2 is* $\frac{1}{4}$ and the probability that they come from a sire is therefore $\frac{1}{4}N_n$(similarly for females). Thus, the probability that two alleles that merge at generation *t* in a zygote come from the same individual (regardless of sex) is $\frac{1}{4}N_n + \frac{1}{4}N_m$.In an idealized population this probability is equal $\frac{1}{N}$,what means that $\frac{1}{N_e}$(Wright, 1938). From this relationship it is then possible to obtain that

$$N_e = \frac{4N_n N_m}{N_n + N_m}. \quad (8)$$

It follows from the above that $N = N_n + N_m, n_s = \frac{N_n}{N}, m = \frac{N_m}{N}$,hense $N_e = 4n_s mN$, which means that N_e is maximal (and corresponds directly to N) when $n_s = m = \frac{1}{2}$. In other cases is $N_e < N$. Furthermore, the above shows that the less numerous genders has a higher impact on the value of N_e. For example, if

$n_s = 0.01$ *i.e..* $N_n = 0.01N$ hence $N_m = 0{,}99N$, hence $N_e \approx 0.0N = 4N_n$.

Population size over generations (Caballero, 1994)

In an idealized population, mating individuals are constant (N, or Nn, Nm) over generations. If we consider the situation where the number of individuals varies over generations with Ni individuals per generation i, the expected heterozygosity in generation t expressed relative to the heterozygosity in the base generation is

$$\frac{H_t}{H_o} = \Pi_{i=1}^{t}\left(1 - \frac{1}{2N_i}\right). \quad (9)$$

If we replace Ne with N in relation (48) and (51) we get

$$\frac{H_t}{H_o} = \left(1 - \frac{1}{2N_e}\right)^t. \quad (10)$$

If the population is large and the number of generations small, this relationship can be adjusted to

$$\frac{1}{N_e} \approx \frac{1}{t}\Sigma_{i=1}^{t}\frac{1}{N_i}\text{(Wright,1938)}. \quad (11)$$

Taking into account the different number of males and females (8), the relation (11) can be adjusted to:

$$\frac{1}{N_e} \approx \frac{1}{t}\Sigma_{i=1}^{t}\left(\frac{1}{4N_{n,i}} + \frac{1}{N_{m,i}}\right). \quad (12)$$

Since this is a harmonic mean, two important points follow from the relationship:

1) Maximum N_e occurs when $\Sigma_{i=1}^{t}N_i$is a constant population size across generations.

2) N_e is significantly affected by the reduction in the number of individuals in one period.

This suggests that the bottleneck effect, which caused an increase in the value of inbreeding, cannot be compensated for by a subsequent increase in population size. This is because the frequency of possible mutations is low even in large populations.

As already mentioned, genetic drift occurs two generations earlier than inbreeding. If the population size changes over generations, genetic drift depends on the number of individuals in a generation. Whereas inbreeding depends on the number of individuals in the grandparents' generation.

Estimation of the effective population size from the actual population

As mentioned, the above procedures are derived from an idealized population. For example, equation (8) represents the possibility of accounting for the different numbers of males and females in the population. In addition, equation (12) accounts for possible fluctuations in the number of individuals from generation to generation. However, these relationships do not consider relatedness between individuals and consider male and female individuals unrelated. However, Cervantes et al. (2008) state that the assumption of an idealized population helps derive the effective population size in a real population in which selection, non-random mating, and overlapping generations occur. Gutiérrez et al. (2003) proposed a procedure applicable in a real population that takes advantage of the increase of inbreeding by year of birth (relationship 7), where $\Delta F = \frac{F_t - F_{t-1}}{1-F_{t-1}}$,where F_t and F_{t-1} the average inbreeding coefficients are in the year (generation) *t* and *t-1*. However, Gutiérrez et al. (2008) report that this method may not be reliable when there is a change in the mating system. For example, if after a period when close relatives were mated, a change occurs and less closely related individuals are mated, this method may provide unrealistic negative values of *Ne*. Gutiérrez et al. (2008) subsequently derived a relationship to estimate the realized effective population size corresponding to the actual populations. This method can be derived as follows:

Assume a population of population size *N* corresponding to the conditions of the idealized population. Under these conditions, the value of the inbreeding coefficient of the hypothetical generation *t* can be obtained from the equation:

$$F_t = 1 - (1 - \Delta F)^t. \tag{13}$$

This idea is based on the inbreeding coefficient values and the equivalent of non-overlapping generations (Maignel et al., 1996) for each individual included in the reference population. Suppose it is further assumed that all individuals in the reference population have the same value of the inbreeding coefficient (F_t=F_i). In that case, the increase in the inbreeding coefficient can be defined as:

$$\Delta F_i = 1 - \sqrt[EqG_i - 1]{1 - F_i}, \tag{14}$$

where *EqGi*, in this case, represents the equivalent of the completeness of ancestral generations (Maignel et al., 1996). The average value of the completeness equivalent of ancestral generations (Maignel et al., 1996) corresponds to the number of generations in discrete (non-overlapping) populations (Wooliams & Mäntysaari, 1995). This implies that the relationship can be applied to real populations. The set of ΔF_i values estimated for each individual in the reference population is then used to estimate the *Ne* of the reference population. The realized *Ne* is then directly derived from the average value of ΔF_i according to the relation:

$$N_{eF} = \frac{1}{2\overline{\Delta F}}. \qquad (15)$$

Another method that corresponds to real populations was proposed by Cervantes et al. (2008). This method is based *on the increase in the value of the coancestry coefficient:* where the increase in the value of the coancestry coefficient (f_{XY}) for all individuals j and k (ΔC_{jk}, Cervantes et al., 2011) is used to realize the effective population size, taking into account $\Delta C_{jk} = 1 - \sqrt[\frac{EqG_j+EqG_k}{2}]{1-C_{jk}}$, where $_{EqGj}$ and $_{EqGk}$ are the equivalent complete generations (Maignel *et al.*, 1996) of individuals j and k, ΔC_{jk} is the increase in the coancestry coefficient between a pair of individuals j and k, and C_{jk} is the inbreeding coefficient of the possible offspring of individuals j and k. The realized effective population size can be estimated using the equation (Cervantes et al., 2011):

$$N_{eC} = \frac{1}{2\overline{\Delta C}}, \qquad (16)$$

where $\bar{\Delta}C$ is the average increase in the coancestry coefficient between individuals j and k in the reference population.

Differences between the estimates of effective population sizes based on the increase in the inbreeding coefficient (N_{eF}) and the increase in the coancestral coefficient between two individuals (N_{eC}) provide information about possible non-random mating between individuals in the population or subpopulations under study and possible reduction in the value of genetic diversity in subsequent generations due to combinations of parental pairs. If the population were considered as an idealized population, N_{eF} and N_{eC} should have identical values. A discrepancy between these parameters indicates a preference for the selection of parental pairs. In other words, comparing N_{eC} and N_{eF} values allows the characterization of the effect of preferential mating in the population (Cervantes et al., 2011). Suppose the values of the N_{eC}/N_{eF} ratio are less than 1. In that case, there is no preferential mating in the population under study, which would split the population into other subpopulations. Conversely, if the N_{eC}/N_{eF} ratio values are greater than 1, it would indicate that preferential mating is occurring in the population, resulting in the population being split into other subpopulations. Within these subpopulations, the relatedness values would be higher than between subpopulations, and therefore, the overall relatedness would be lower than the average inbreeding coefficient.

The realized effective population size can be interpreted as the total effective size over time that led to the current level of the inbreeding coefficient from the founder population.

References

Caballero, A. 1994. Developmnets in the prediction of effective population-size. Heredity. 73 (6). 657-679.

Cervantes, I., Goyache, F., Molina, A., Valera, M., Gutiérrez, J. P. 2011. Application of individual increase in inbreeding to estimate effective size from real pedigrees. Journal of Animal Breeding and Genetics. 125. 301–310.

Daetwyler, H.D, Villanueva, B., Bijma, P., Woolliams, J.A. 2007. Inbreeding in genome-wide selection. Journal of Animal Breeding and Genetics 124:369–376.

Dagnachew, B.S, Meuwissen, T.H.E. 2016. A fast Newton–Raphson based iterative algorithm for large scale optimal contribution selection. Genetics Selection Evolution 48.

Dalvit C, de Marchi M, Zanetti E, Cassandro M. 2009. Genetic variation and population structure of Italian native sheep breeds undergoing in situ conservation1. Journal of Animal Science 87:3837–3844.

Ellegren, H., Galtier, N. 2016. Determinants of genetic diversity. Nature Reviews Genetics 17:422–433

Falconer, D. S., Mackay, T. F. C. 1996. Introduction into quantitative genetics. Longman House, Harlow Essex., p. 464. ISBN: 978-0582243026.

FAO, 2000. Secondary guidelines for development of farm animal genetic resources management plans. Management of small populations at risk. FAO. Rome. Italy. p. 219.

Forutan M, Ansari Mahyari S, Baes C, Melzer N, Schenkel FS, Sargolzaei M. 2018. Inbreeding and runs of homozygosity before and after genomic selection in North American Holstein cattle. BMC Genomics 19:1–12.

Gutiérrez, J. P., Cervantes, I., Molina, A., Valera, M., & Goyache, F. (2008). Individual increase in inbreeding allows estimating effective sizes from pedigrees. *Genetics selection evolution*, *40*(4), 359-378.

Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across mammalian genomes. Nature Reviews Genetics 12:756–766.

Lynch M. 2010. Evolution of the mutation rate. Trends in Genetics 26:345–352.

Maignel, L., Boichard, D., Verrier, E. 1996. Genetic variability of French dairy breeds estimated from pedigree information. Interbull Bull, 14. 49-54.

Makanjuola, B.O, Miglior, F., Abdalla, E.A., Maltecca, C., Schenkel, F.S., Baes, C.F. 2020. Effect of genomic selection on rate of inbreeding and coancestry and effective population size of Holstein and Jersey cattle populations. Journal of Dairy Science 103:5183–5199.

Mendelsohn, R., 2003. The challenge of conserving indigenous domesticated animals. Ecological Economics 45:501–510.

Mészáros G. 2018. Genomic descriptors of biodiversity – A review. Die Bodenkultur: Journal of Land Management, Food and Environment 69:73–83. Available from https://www.sciendo.com/article/10.2478/boku-2018-0007.

Notter DR. 1999. The importance of genetic diversity in livestock populations of the future. Journal of Animal Science 77:61. Available from https://academic.oup.com/jas/article/77/1/61-69/4625321.

Relichova, J. 2009. Genetika populací. Masarykova univerzita, ISBN: 978-80-210-4795-2

Saravanan KA, Panigrahi M, Kumar H, Bhushan B. 2022. Advanced software programs for the analysis of genetic diversity in livestock genomics: a mini review. Biological Rhythm Research 53:358–368.

Scott, B.A, Haile-Mariam, M., Cocks, B.G., Pryce, J.E. 2021. How genomic selection has increased rates of genetic gain and inbreeding in the Australian national herd, genomic information nucleus, and bulls. Journal of Dairy Science 104:11832–11849.

Sonesson, A.K, Woolliams, J.A, Meuwissen, T.H. 2012. Genomic selection requires genomic control of inbreeding. Genetics Selection Evolution 44.

Tapio M, Marzanov N, Ozerov M, Ćinkulov M, Gonzarenko G, Kiselyova T, Murawski M, Viinalass H, Kantanen J. 2006b. Sheep Mitochondrial DNA Variation in European, Caucasian, and Central Asian Areas. Molecular Biology and Evolution 23:1776–1783.

van Wyk JB, Fair MD, Cloete SWP. 2009. Case study: The effect of inbreeding on the production and reproduction traits in the Elsenburg Dormer sheep stud. Livestock Science 120:218–224.

Vicente, A. A., Carolin, N., Gama, L. T. 2014. Genetic diversity in th Lusitano horse breed assessed by pedigree analysis. Livestock Science. 148. 16-25.

Vignal A, Milan D, SanCristobal M, Eggen A. 2002. A review on SNP and other types of molecular markers and their use in animal genetics. Genetics Selection Evolution 34:275–305.

Wright, S. 1968. The Theory of Gene Frequencies: Evolution and the Genetics of Populations, Vol. 2. Chicago University press, Chicago, USA. p. 520. ISBN: 9780226910390.

Wright, S. Size of population and breeding structure in relation to evolution. Science 87, 430–431 (1938).

Wooliams J.A., Mäntysaari E.A. 1995. Genetic contributions of Finnish Ayrshire bulls over four generations. Animal Science, 61: 177-187.

Woolliams, J.A., Berg, P., Dagnachew, B.S., Meuwissen, T.H.E. 2015. Genetic contributions and their optimization. Journal of Animal Breeding and Genetics 132:89–99.