Using the Information Collected for Genetic Evaluation to Assess the French Ruminant and Equine Breeds' Genetic Variability

C. Danchin-Burge[1], L. François[1], D. Laloë[2], G. Leroy[2] and E. Verrier[2]
[1] *Institut de l'Elevage, 149, rue de Bercy, 75012 Paris, France*
[2] *INRA/AgroPariTech, UMR1313 GABI, 78350, Jouy-en-Josas, France*
coralie.danchin@idele.fr

Abstract

In livestock species, selection programs are getting more and more efficient. Meanwhile, the development of new tools such as genomic selection will probably fasten how breeds are evolving. However, preserving the breeds' genetic variability is a necessity. First of all, genetic progress is correlated with this factor. Also, a sharp increase in inbreeding in a breed might have a very negative impact on production and functional traits. Therefore, in order to achieve the sustainability of our selection choices for the future, the breeds' genetic variability requires an accurate management. To do so, two type or data can be used to calculate variability genetic indicators: pedigrees and genotype data. These indicators can be used to monitor the impact of a main ancestor in a breed for instance. The VARUME (Genetic VARiability of Ruminants and Equids) project goal is to set up a genetic variability observatory for the Ruminants and Equids species. It will generate indicators that can assess breed genetic variability on a regular basis, by using a common method that will be acknowledged and used by all breed managers.

Key words: Genetic variability, Monitoring, Ruminants, Equids, Pedigree, Molecular information

Introduction

In France, selection programs in ruminant species are extremely efficient and are a major contributor to the proficiency of the meat and dairy industries. Meanwhile, the development of new tools such as genomic selection will probably fasten how breeds are evolving. The downside is a loss of genetic variability in most breeds which is documented worldwide (see for instance, for cattle: Mc Parland *et al.*, 2007; Danchin-Burge *et al.*, 2011). Preserving the breeds' genetic variability is a necessity. First of all, genetic progress is correlated with this factor. Also, a sharp increase in inbreeding in a breed might have very negative impacts through inbreeding depression (see for instance Croquet *et al.*, 2006) or higher frequencies of genetic defects (such as CVM, Agerholm *et al.*, 2001). There is therefore a need to provide genetic variability indicators, on a regular basis, so that breeds' genetic managers can adjust their management accordingly.

Several types of information and different methods are used to assess the genetic variability of a population (see for instance Baumung and Sölkner, 2002; Valera *et al.*, 2005; Engelsma *et al.*, 2010). Generally, two types of data are used to calculate variability genetic indicators: pedigrees and genotype data. In France, pedigree information is widely used to evaluate the breeds' genetic variability. However, studies per breed were mostly conducted on a one time basis. Also, with the setting up of genomic selection, in dairy cattle and sheep, there are now numerous molecular data that are generated for the needs of selection programs. These markers also represent an interesting source of information to create genetic variability indicators.

In our article, after briefly presenting the general aim of the VARUME project, we will describe how various indicators based on pedigrees data were assessed and chosen to be part of the VARUME observatory. We will

finish by presenting the first results of an ongoing study on how to choose the best method to calculated the effective population based on molecular data.

The VARUME project

The main organizations involved in the French breeding programs collaborated in order to create a project called VARUME (for genetic VARiability of Ruminants and Equids). The general aim of this project is to build a genetic variability observatory of all the ruminants breeds (i.e. dairy cattle, sheep and goat, meat cattle and sheep) and equids breeds (i.e. horses and donkeys) that are under selection or conservation in France. One part of the project is to define a list of indicators that are globally acknowledged and understood by the breeds' managers. In the end, genetic variability indicators will be provided on a regular basis to all the organizations in charge of the management of a breed. For most breeds, these indicators will also be available on line for anyone interested by the topic, and regular training will be provided to make sure that all indicators are fully understood by the people that are going to use them.

Observatory based on pedigrees

Choice of software

Various software can be used to calculate genetic variability indicators based on pedigrees. PEDIG (Boichard, 2002) and ENDOG (Gutiérrez and Goyache, 2005) are probably the most quoted in the literature. ENDOG is more user friendly than PEDIG but the latter was chosen since it can handle millions of pedigree data and doesn't block data calculations if an offspring is born the same year as its parents, which is not the case with ENDOG.

Choice of populations under study

A first task of the VARUME partners was to define what would be the populations under the study. It was agreed that indicators would be published per breed, the signification of a breed being the one agreed on by the common regulations in force. In each breed, the indicators will be calculated for a given "analyzed population" (P_{an}). Within breed, analyzed population were defined as, for the ruminant species, (1) all the females born within the last four years (i.e. the rough equivalent of a generation interval) (2) all the AI males that sired at least one of the female part of the previous analyzed population. The definition is slightly different for equids since the analyzed populations are all the animals born the last four years, including males, gelding and females. The choice for equids was done since it is very difficult to define a breeder population in these species: for instance females might just used for leisure, males could be gelded without notification to the national database, and some geldings sired offspring before being gelded…

Any breed acknowledged by national regulation will have indicators published as long as sufficient pedigree information exists for the breed. "Sufficient information" was defined as an average of 2.5 generations known for the female analyzed population. This number is empiric and based on the data obtained from previous studies.

Pedigree information

For the ruminants species, all the data are coming from the National Genetic Information System (French acronym: SNIG) that are stored on servers from the Genetic Information Data Centre of INRA. For the equids species, the information are coming from the Equids Information System (SIRE) managed by IFCE

(Institut Français du Cheval et de l'Equitation). All the pedigree information are collected and used mainly for genetic evaluation, therefore their use for the creation of genetic variability indicators doesn't require any additional costs.

Observatory based on SNP data

Till recently, the main limit for the use of molecular data to characterize within breed diversity was a limited number of molecular markers. Thanks to genomic selection, it is now possible to have a very fine characterization thanks to the 50 k or even 800 k SNP chips, at a reasonable cost. In France, numerous projects are using these chips for genomic selection, finding new QTL or selection signatures. These markers are also valuable source of information to characterize genetic variability, however there are not very often used to do so. Most of the works on genetic diversity are focalized primarily on species and/breeds diversity (Groeneveld *et al.*, 2010 ; in cattle : The Bovine HapMap Consortium *et al.*, 2009 et Laloë *et al.*, 2010; in sheep : Kijas *et al.*, 2009).

The second work package of the VARUME project focuses on the feasibility of creating genetic variability indicators based on available SNP data. The genotypes provided by various consortiums (cf. acknowledgment) are coming from 3 dairy cattle breeds and 4 sheep dairy breeds. Other genotypes exist in other industries such as meat cattle and sheep, however we were expecting to have sampling effect problems. Indeed, in the dairy industry, almost all the males are used through AI and genotyped, which is not the case in the meat industry where most of the breeding is done with natural service males.

Assessment of different indicators based on pedigrees

In this part we will present what indicators were chosen for the VARUME projects. All these indicators are widely known and described in various articles (such as Maignel *et al.*, 1996), therefore a sole simple definition will be provided for each of them.

Quality control

Breeds managers are likely to use the indicators to orientate their breeding program, therefore they need accurate indicators. One of the main hypotheses used to build most pedigree indicators is that founders (i.e. animals without pedigree information) are not related. It is a strong postulate which is erroneous most of the time, however its impact is limited if most founders were born several generations before the analyzed populations.

Therefore, one way to judge the accuracy of the indicators based on pedigrees is to estimate its pedigree depth. This can be obtained for instance by an indicator called equivalent number of know generations (Eq) which is obtained by summing the proportion of ancestors *i* known at each generation *n*. For instance, a breed with an analyzed population with an Eq of 5 is a breed that has on average, for each animal from the analyzed population, 5 generations of ancestors known. According to Baumung and Sölkner (2002), an average of 5 Eq is sufficient to obtain reliable inbreeding indicators.

Demographic indicators

The main advantage of demographic indicators is that they are easy to grasp and therefore very valuable to make breeds managers understand a point. They are also useful for breeds with little pedigree information. Indicators can be static and calculated for an analyzed population (number of sires, dams, average number of offspring per sire, maximum number of offspring etc.) or show the evolution of the breed over a longer period (total number of females over a 10 year period etc.). For the analyzed population, it is always useful to have a comparison between the total number of females (F_{tot}) and the number of females with 2 parents known (F_{an}). F_{an} is used to calculate most genetic variability indicators, and in some breeds, mostly the hardy ones, the percentage of unknown sire can be high. Therefore the genetic variability analysis relies on a small percentage of the total population and it is necessary to have a way to advertise this to breeds managers.

Indicators based on probability of origin

The methods based on the probability of gene origin are a simulation from which animals are coming the genes existing in an analyzed population. These indicators are calculated in order to find who the major ancestors of the population are, i.e. the main common ancestors between the different individuals. In order to calculate the influence of each ancestor, some postulates are made:

- The only source of the diversity are the founders, which are considered as unrelated (i.e. mutations are not taken into account).

- For a given individual, the probability for an allele to be inherited from one of its parents its ½, ¼ for one of its grand-parents and so forth.

The probability to have transmitted an allele to an individual of the P_{an} is calculated for each ancestor of the $_{Pan.}$ The expected contribution of an ancestor to the genome of an individual is called probability of gene origin. The contribution of each ancestor to the P_{an} is calculated by summing the probabilities. At first, the contributions are calculated without taking into account how the different ancestors are related: there are called raw contributions.

In a second step, only the marginal contribution of an ancestor, i.e., after ranking ancestors by decreasing contribution, the contribution not yet explained by the previous ancestors, are calculated. An effective number of founders Fe is then calculated based on the raw contributions of the founders and an effective number of ancestors Ae is calculated with the marginal contributions of the ancestors. The Ae is equal to the number of ancestors that will be needed, if they all had equal contributions, to generate the same level of genetic variability than the analyzed population. Systematically Fe is superior to Ae. The Ae/Fe ratio is a way to detect bottlenecks events in a population. For instance, with equivalent Ae number, on a general basis a selected breed will have a smaller ratio than a rare breed.

One of the main advantage of these indicators is that they are not as dependant as inbreeding (see after) of pedigree depth. They are "historical" indicators, i.e. they help us tracing what were the main events in the history of the genetic variability of the breed.

Another way to use the gene probability of origin is to trace the use of exogenous gene of a breed. To do so, each founder is attributed a different origin (5 different origins is generally a maximum to obtain clear results) and the evolution of their genes is followed from one generation to the other, by using the gene probability of origin.

Indicators based on probability of identities

The underlying question for these indicators is if two alleles carried by an animal are identical. Two animals are said to be related if they have common ancestor, and an animal is inbred if his parents are related. Two alleles are said to identical by descent if they are the duplication of a same ancestral allele. The inbreeding coefficient of an individual is therefore the probability that two alleles carried by an individual are identical by descent, and it is equal to the kinship coefficient between its parents.

The consequence of inbreeding raise is increasing the breeds' genome homogenization. Therefore calculating an inbreeding rate (i.e. inbreeding increase over time) is a good indicator of a loss of genetic variability. An inbreeding coefficient by itself is not a good indicator since it is linked to pedigree knowledge: inbreeding level always increases with pedigree depth.

Another common indicator is the effective population size (Ne) but we will talk in details about this indicator in the following paragraphs.

Assessment of different indicators based on genotyped data

The main advantage of calculating an effective number of alleles is its easy interpretation: when there is a 10% loss in diversity, 10% of the alleles are lost in a population (if equally frequent). However, it gives no information about the number of alleles present at a certain locus (Allendorf *et al.,* 2013; Jost, 2008).

The proportion of polymorphic loci, also known as polymorphism, is the mean number of heterozygous loci. It gives an indication of the percentage of polymorphic loci in a population and can be used with codominant markers such as SNPs. However, this method is very dependent on sample size as it is more likely to detect genetic variation when more individuals are genotyped. To avoid this dependency, a limit is set to the frequency of the most common allele, usually at 0,99 or 0,95 (Allendorf *et al.,* 2013; Hedrick, 2005).

The effective population size (Ne) is the size of an idealised population which could give rise to the rate of inbreeding or the rate of change in variance of gene frequencies observed in the population under consideration (Wright, 1931).

There are different ways of calculating the current effective population size based on SNP data. We will focus here on the linkage disequilibrium between physically unlinked loci to estimate Ne. This method was first developed by Weir & Hill (1980) for both linked and unlinked loci and adapted to be used solely for unlinked loci (Waples, 1991). Our data were generated from the genotypes used for genomic selection in 4 dairy sheep breeds, respectively the Lacaune, Basco-Béarnaise (BB), Manech Tête Rousse (MTR) and Manech Tête Noire (MTN).

Obtaining an accurate value for Ne

Based on SNP data obtained for 4 dairy sheep breeds, various scenario cases were done in order to see what the minimum number of samples were needed to obtain an accurate value for Ne. The formulae used to calculate Ne was the one given by Waples (2006), where r^2 is the allelic correlation and *n* the number of samples. Waples suggest using a sampling correction of 1/*n* instead of 1/2*n:*

$$Ne = \frac{1}{3(r^2 - \frac{1}{n})}$$

In a study of England *et al.* (2006), it was found that if the true Ne was larger than the sample size used to estimate it, there was a substantial bias on the estimation of Ne using the linkage disequilibrium method. The bias was not affected by the number of loci but could be alleviated by sampling more individuals. For the detection of possible bias it suggested that subsampling of the existing sample up to the full sample size could be used. From the shape of the curve of sample size versus Ne, it can be seen whether or not bias was present. Waples (2006) also described this bias caused by a small sample size and suggested that the second order terms should not be ignored when the sample size is small.

It was examined if the proposed correction of 1/n was sufficient to account for the effect of sample size. Looking at Figure 1 for the BB breed, we can see that the correction works well for sample sizes larger than 60 while it is not sufficient for smaller sample sizes where there is a clear bias in the estimation of Ne.

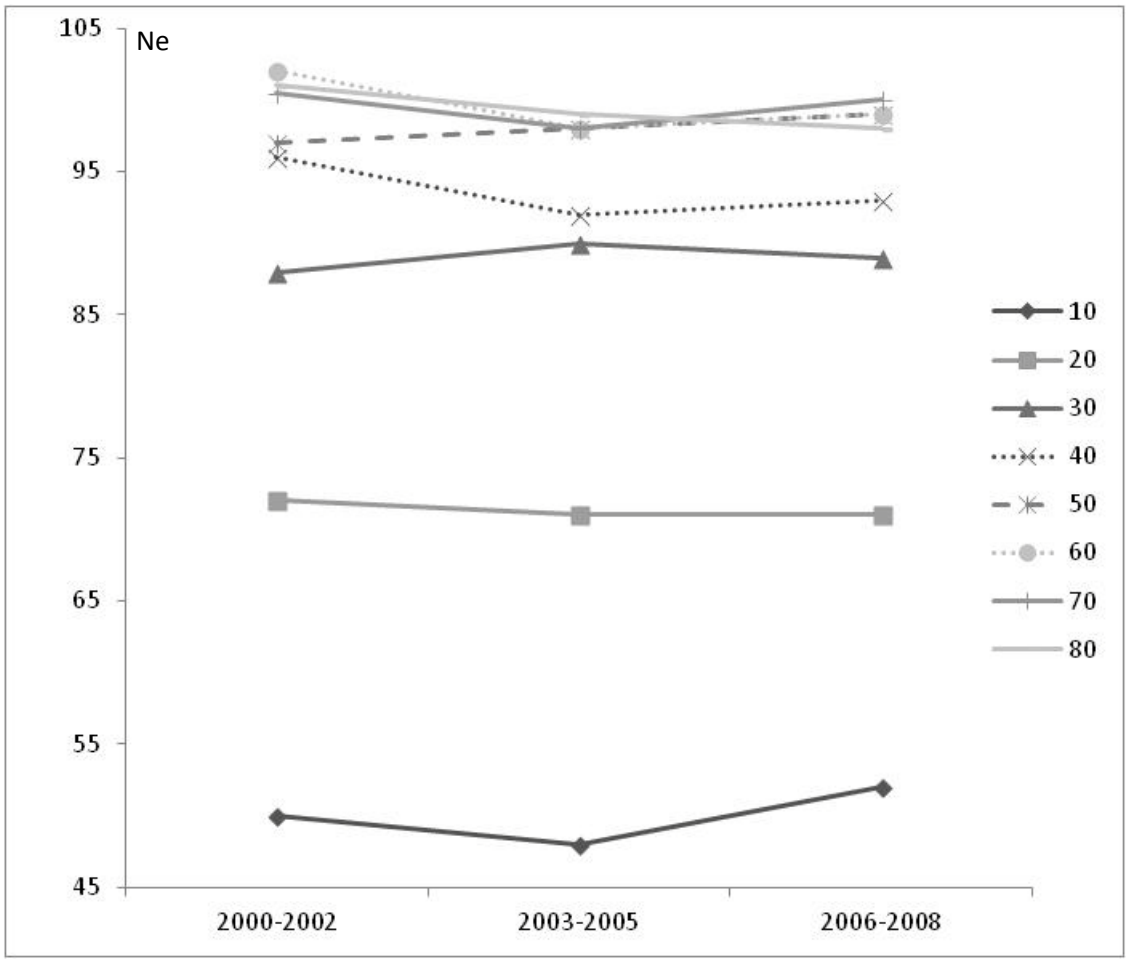

Figure 1. Ne over generation and by sample size for the Basco-Béarnaise breed.

Both for the Lacaune (results not shown) and Basco-Béarnaise breeds, the Ne stabilises for a sample size between 60 and 80 samples. This sample size seems to be the minimum to avoid bias.

Furthermore, Waples (2006) suggested that for smaller sample sizes, the second order term may not be ignored and in his study the formula was adapted empirically. In our study, we found that the second order term gave no significant improvement for smaller sample sizes. We can conclude that this method works well if the samples have an appropriate size.

Effect of substructure on Ne

Because of selection and relatedness between animals, a correlation between alleles on different chromosomes is created. Therefore an allelic correlation (r^2) can be calculated between each pair of alleles of two unlinked loci and by averaging the r^2 values over all pairs of alleles to give one single value for each pair of loci.

Figure 2. PCA analysis based on molecular kinship coefficients in the Lacaune breed.

When structure is present in a breed (see Figure 1 for instance for the Lacaune breed) an artificial correlation may be created due to the differences in allele frequencies in the subpopulation. It is possible to account for the effect of structure using the linkage disequilibrium between unlinked loci. When one does not account for structure, the r^2 value may be biased because of the effect of structure. While the two subpopulations might only have a small but different r^2 between two loci, the method finds r^2 value that is higher since it takes the mean of the whole population, giving a lower estimation of Ne due to their inverse relationship. Assigning individuals to their proper subpopulations, the method will compute the partial correlation between markers, given the structure. The r^2 will be reduced and so the estimated Ne will increase (Mangin *et al.*, 2011).

Comparison with pedigree data

To validate this molecular method using linkage disequilibrium between unlinked loci, we will compare it with two pedigree methods: the pedigree method from Gutiérrez *et al.* (2009) – Ne Ped F - using the individual increase in inbreeding; and the method from Cervantes *et al.* (2011) using the individual increase in coancestry – Ne Ped ϕ.

The method of Gutiérrez *et al.* (2009) could be considered to have similarities to the method of linkage disequilibrium (Ne LD), as it does not take into account population structure (Leroy *et al.*, 2013), while the method of Cervantes *et al.* (2011) could be corresponding to the method of linkage disequilibrium when one accounts for substructure (Ne LD struct). For the estimation of Ne using these pedigree-based methods, the genotyped individuals were used as reference population. The results are shown in Table 1.

Table 1. Ne values for 4 dairy sheep breeds averaged the 2000-2011 period (four generations) with four different methods.

Ne	Ne LD	Ne LD struct	Ne Ped F	Ne Ped ϕ
Lacaune	195	303	223	312
MTR	118	145	153	148
BB	98	/	108	91
MTN	92	/	82	82

When taking the average over the last four generations, the molecular methods correspond well to the pedigree methods in general (Table 1).

Final discussion on Ne

To use the estimation of the effective population size based on linkage disequilibrium between unlinked loci to estimate the genetic variability in populations, we found that the number of genotyped individuals needs to be greater than sixty. If there are less individuals genotyped the method is biased as was also found in previous studies (England *et al.*, 2006) and (Waples, 2006). The genotyped population should be representative of the total population and over generation it should be constituted of the same kind of animals (if only males are genotyped in the first generations, this should not be changed in the latter generations).

Conclusion

There are strong demands coming from rare breeds' managers as well as selected breeds organizations to have at their disposals genetic variability indicators, on a regular basis. These needs are increased with genomic selection since generation intervals are shortened. It becomes crucial to be able to evaluate quickly the impact of the new selection programs on the breeds' genetic variability in order to maintain their sustainability.

The VARUME project should insure a secure way to monitor each breed's genetic variability. It is also a way for France to consolidate its commitments toward the protection of its genetic resources, following other ongoing actions such as the French National Cryobank (created in 1999) or the following of conservation programs by the Institut de l'Elevage (ruminant species) and IFCE (equid species).

Acknowledgment

Our deep thanks to Roquefort'IN and GENOMIA for providing us the dairy sheep

genotypes. Funding provided by CASDAR (French Minister of Agriculture).

References

Agerholm, J.S., Bendixen, C., Andersen, O., & Arnbjerg, J. 2001. Complex vertebral malformation in Holstein calves. *J. Vet. Diagn. Invest. 13,* 283–289.

Allendorf, F., Luikart, G. & Aitken, S. 2013. *Conservation and the Genetics of Populations.* Wiley-Blackwell, Chichester, UK, 2nd edition. 602 p.

Baumung, R. & Sölkner, J. 2003. Pedigree and marker information requirements to monitor genetic variability. *Geneti. Sel. Evol. 35,* 369-383.

Boichard, D. 2002. PEDIG: a Fortran package for pedigree analysis suited for large populations. *Proc. of the 7th WCGALP*, Montpellier, France. Communication No. 28-13.

Cervantes, I., Goyache, F., Molina, A., Valera, M. & Gutiérrez, J.P. 2011. Estimation of effective population size from the rate of coancestry in pedigreed populations. *J. Anim. Breed. Genet. 128,* 56-63.

Croquet, C., Mayeres, P., Gillon, A., Vanderick, S. & Gengler, N. 2006. Inbreeding depression for global and partial economic indexes, production, type and functional traits. *J. Dairy Sci. 89,* 2257–2267.

Danchin-Burge, C., Leroy, G., Brochard, M., Moureaux, S. & Verrier, E. 2011. Evolution of the genetic variability of eight French dairy cattle breeds assessed by pedigree analysis. *J. Anim. Breed. Genet. 129*, 206–217.

Engelsma, K.A., Veerkamp, R.F., Calus, M.P. & Windig, J.J. 2011. Consequences for diversity when prioritizing animals for conservation with pedigree or genomic information. *J. Anim. Breed. Genet. 128:6,* 473-81.

England, P.R., Cornuet, J.-M., Berthier, P., Tallmon, D.A. & Luikart, G. 2006. Estimating effective population size from linkage disequilibrium: severe bias in small samples. *Cons. Genet. 7:2,* 303-308.

Groeneveld, L.F., Lenstra, J.A., Eding, H., Toro, M.A., Scherf, B., Pilling, D., Negrini, R., Finlay, E.K., Jianlin, H., Groeneveld, E., Weigend, S. & The GLOBALDIV Consortium, 2010. Genetic diversity in farm animals – a review. *Anim. Genet. 41 (Suppl. 1),* 6–31.

Gutiérrez, J.P. & Goyache, F. 2005. ENDOG: a computer program for monitoring genetic variability of populations using pedigree information. *J. Anim. Breed. Genet. 122,* 357-360.

Gutiérrez, J.P., Cervantes, I. & Goyache, F. 2009. Improving the estimation of realized effective population sizes in farm animals. *J. Anim. Breed. Genet. 126:4,* 327-332.

Hedrick, P. 2005. *Genetics of populations.* Jones and Bartlett Publishers, Sudbury, Massachusetts, 3rd edition. 737 p.

Jost, L. 2008. GST and its relatives do not measure differentiation. *Mol. Ecol. 17:18,* 4015-4026.

Kijas, J.W., Townley, D., Dalrymple, B.P., Heaton, M.P., Maddox, J.F., McGrath, A., Wilson, P., Ingersoll, R.G., McCulloch, R., McWilliam, S., Tang, D., McEwan, J., Cockett, N., Oddy, V.H., Nicholas, F.W. & Raadsma, H. 2009. A Genome Wide Survey of SNP Variation Reveals the Genetic Structure of Sheep Breeds. *PLoS ONE* 4(3): e4668. doi:10.1371/journal.pone.0004668.

Laloë, D., Moazami-Goudarzi, K., Lenstra, J.A., Marsan, P.A., Azor, P., Baumung, R., Bradley, D.G., Bruford, M.W., Cañón, J., Dolf, G., Dunner, S., Erhardt, G., Hewitt, G., Kantanen, J., Obexer-Ruff, G., Olsaker, I., Rodellar, C., Valentini, A., Wiener, P. & European Cattle Genetic Diversity Consortium and Econogene Consortium. 2010. Spatial Trends of Genetic Variation of Domestic Ruminants in Europe. *Diversity 2:6,* 932–945.

Leroy, G., Mary-Huard, T., Verrier, E., Danvy, S., Charvolin, E. & Danchin-Burge, C. 2013. Estimating the effective population size using pedigree data. What method should be used in practice? Examples in dogs, sheep, cattle and horses. *Genet. Sel. Evol. 45,* 1. doi:10.1186/1297-9686-45-1.

Maignel, L., Boichard, D. & Verrier, E. 1996. Genetic variability of French dairy breeds estimated from pedigree information. *Interbull Bulletin 14,* 49–54.

Mangin, B., Siberchicot, A., Nicolas, S., Doligez, A., This, P. & Cierco-Ayrolles, C. 2011. Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity 108:3,* 285-291.

Mc Parland, S., Kearney, J.F., Rath, M. & Berry, D.P. 2007. Inbreeding trends and pedigree analysis of Irish dairy and beef cattle populations. *J. Anim. Sci. 85,* 322–331.

The Bovine HapMap Consortium, Gibbs, R.A., Taylor, J.F., Van Tassell, C.P., Barendse, W., Eversole, K.A., Gill, C.A., Green, R.D., Hamernik, D.L., Kappes, S.M. *et al.*, 2009. Genome-Wide Survey of SNP Variation Uncovers the Genetic Structure of Cattle Breeds. *Science 324,* 528-532.

Valera, M., Molina, A., Gutiérrez, J.P., Gómez, J. & Goyache, F. 2005. Pedigree analysis in the Andalusian horse: population structure, genetic variability and influence of the Carthusian strain *Livest. Prod. Sci. 95,* 57–66.

Waples, R. 1991. Genetic methods for estimating the effective size of cetacean populations. *Genetic Ecology of Whales and Dolphins*, pp. 279-300.

Waples, R. 2006. A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. *Cons. Genet. 7,* 167-184.

Weir, B.S. & Hill, W.G. 1980. Effect of mating structure on variation in linkage disequilibrium. *Genetics 95:2,* 477-488.

Wright, S. 1931. Evolution in Mendelian populations. *Genetics 16:2,* 97-159.