Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

POPREP: a generic report for population management

E. Groeneveld[1], B.v.d. Westhuizen[2], A. Maiwashe[2], F. Voordewind[2] and J.B.S. Ferraz[3]

[1]Friedrich Loeffler Institute, Institute of Farm Animal Genetics, Department of Animal Breeding and Genetic Resources, Neustadt, Germany
[2]Agricultural Research Council, Animal Production Institute, Animal Breeding and Genetics Unit, Irene, Republic of South Africa
[3]Departamento de Ciências Básicas, Grupo de Melhoramento Animal e Biotecnologia, Faculdade de Zootecnia e Engenharia de Alimentos, Universidade de São Paulo, Pirassununga, SP, Brasil

Corresponding author: E. Groeneveld
E-mail: eildert.groeneveld@fli.bund.de

Genet. Mol. Res. 8 (3): 1158-1178 (2009)
Received July 3, 2009
Accepted July 31, 2009
Published September 29, 2009

ABSTRACT. Genetic variation provides a basis upon which populations can be genetically improved. Management of animal genetic resources in order to minimize loss of genetic diversity both within and across breeds has recently received attention at different levels, e.g., breed, national and international levels. A major need for sustainable improvement and conservation programs is accurate estimates of population parameters, such as rate of inbreeding and effective population size. A software system (POPREP) is presented that automatically generates a typeset report. Key parameters for population management, such as age structure, generation interval, variance in family size, rate of inbreeding, and effective population size form the core part of this report. The report includes a default text that describes definition, computation and meaning of the various parameters. The report is summarized in two pdf files, named Population Structure and Pedigree Analysis Reports. In addition, results (e.g., individual inbreeding coefficients, rate of inbreeding and effective population size) are stored in comma-separate-values files that are available for

further processing. Pedigree data from eight livestock breeds from different species and countries were used to describe the potential of POPREP and to highlight areas for further research.

Key words: Software; Inbreeding; Effective population size; Cattle; Pigs; Sheep

INTRODUCTION

Conservation of breeds under threat of extinction has long been an issue of importance. In the wake of the Rio Convention (United Nations, 1992), the focus of conservation has expanded to within breed diversity of both small and large populations, which are consequently not under threat of extinction. The reasons for concern lie in the high use of a few genetically superior males, which is made possible by artificial insemination, resulting in increased rates of inbreeding. The Holstein breed, for example, in terms of absolute numbers is one of the most abundant cattle breeds, but it may be losing genetic diversity at an undesirable rate, as indicated by its rather low effective population size (Maignel et al., 1996). While management of biodiversity within small populations is of importance, it is even more relevant in large mainstream populations. As they are responsible for much of our food production, their deterioration would have a widespread negative effect. Therefore, the National Action Program on Animal Genetic Resources in Germany (BMVEL, 2004) stipulates continuous monitoring of active breeding populations. Accordingly, management of populations that are not under direct threat of extinction has also become an issue.

Often, good information on generation intervals, inbreeding rates and effective population sizes, which are required for population management, are not available. In cooperative breeding programs, selection decisions are usually made by individual breeders. The degree of utilization of individual sires results from a sum of individual breeders' decisions. Consequently, the cumulative effect of individual selection decisions can only be assessed *ex post*. However, if trends were detected early on, counteractive measures could still be taken.

We developed a generic report program, based on a minimum set of individual animal information. The intended use is three-fold: firstly, it can serve as documentation of parameters relevant to biodiversity issues to be included in the yearly reporting on populations, as proposed in Groeneveld (2003). Secondly, its outputs can serve as an early warning system in the management of big and small populations, so that negative trends counteracted as soon as they become apparent. Finally, intermediate output can be made available for further research.

MATERIAL AND METHODS

Current animal recording systems and testing schemes facilitate continuous collection of nearly all events in an animal's life. Starting with a birth record, information on performance records is usually also available. Genetic evaluation leads to selection decisions that may be recorded explicitly or can be assumed on the basis of reproduction records. Depending on the frequency of reporting to the central data repository, the statistics generated can be up to date, reflecting current status or be somewhat lagging in up-to-dateness. This set of data allows breeders to create a number of reports that reflect the dynamics of the population. Analyses that make use of this type of data are grouped together in a "Population Structure Report".

With controlled mating and complete reporting, full pedigrees are usually available in modern breeding populations. This allows analysis of the genetic structure of populations that can be grouped in the "Pedigree Analysis Report".

POPREP Data Inputs

Central data repositories are of paramount importance in modern animal breeding and management. While data are normally collected at many different locations, such as herds, test stations, milk labs, and abattoirs, their use normally requires an integrated view of all data. State of the art BLUP genetic evaluation of breeding populations is based on this integration of all performance and pedigree data from the population.

APIIS as a data source

With the development of the Adaptable Platform Independent Information System (APIIS) in animal husbandry, a generalized framework is available to implement such centralized data repositories for any database that needs to store records on individual animals (Groeneveld, 2004).

The unified structure of the databases allows us to develop and run software that is largely independent of the species that the database holds. It is in this context that the "Population Structure Report" and "Pedigree Analysis Report" were developed.

External data sources

As both the population structure and the pedigree analysis reports are extensive, breeding programs that do not stored their data in an APIIS database may also be interested in using them. The minimum requirements are: unique identification of all animals; for each animal the sire, the dam, birthdate, and sex need to be known. The users need to supply these data in a consistent ASCII format, which will then be loaded into an APIIS-conforming database for evaluation and report generation.

Depending on the data collection and storage scheme, pedigree data may comprise different sets of animals. Some recording schemes store records on each animal born; accordingly, the pedigree dataset may contain all animals born and recorded in the breeding program. This would result in a pedigree dataset with roughly equal numbers of males and females. Alternatively, some programs may record only the breeding stock, i.e., those animals chosen to become parents of the next generation. In that case, many more females than males may be in the pedigree file.

Population Structure

Management of a breeding population implies control of matings, a condition that may be more easily met in centrally organized breeding programs than in many cooperatively organized systems. In the former, parameters such as number of males used, their frequency and duration of use are usually well defined and enforced. In the latter, this may not be the case: the sum of individual breeders' decisions determines the population structure, with the result that the composition of the new generation can only be determined *ex post*.

In either case, the number of breeding animals used over time determines the dynamics of a population, which can be described by counts broken down by year. These statistics are well suited for population management purposes. Hence, the basic layout of the tables lists the changes in parameters over the years.

Here, the group of animals or records that give rise to the statistics reported in the tables for one year is called a cohort. In some cases, their definition is obvious, while in others it requires more elaboration. This is why each description of the reports will contain a formal definition of the cohort.

In the Population Structure Report, five sets of tables and figures are generated.

Set 1: Number of breeding males and females

The number of breeding animals at a given time determines the genetic structure of the population in subsequent generations. They determine to a large degree the effective population size N_e, a central parameter in population management.

The following statistics are computed on a per year basis for breeding males and females separately: number of males and females represented in the services (which will be considered automatically if they exist in an APIIS structure) and births, the number of selected animals born in that year and the total number of animals born.

Cohort: *The cohort is the group of selected animals born in a given year. A selected animal is defined as an animal that has participated in a first service, if service dates are available in the dataset, or has already become a parent as indicated by an animal record with the animal in question as a parent.*

Thus, only animals with their data are represented in the cohort; these have not only been selected to become parents, but have either started reproduction through a service/mating or already have offspring in the data repository. This may contrast with many statistics where the number of breeding animals is derived from animals that have been selected for breeding purposes, e.g., total number of cows and bulls in the herdbook. The latter statistics will overestimate the number of active breeding stock, as it includes animals that have never reproduced and also those that might already have been culled but not reported.

Set 2: Age structure of parents

The rate of genetic progress in the population depends, among other things, on the turnover of breeding stock. Thus, the distribution of dams and sires over age classes will be informative in this regard. The tables give an overview of the age structure of the animals' parents per birth year, separately for males and females.

Cohort: *The cohort is defined as the total number of animals born in a given year.* The total number of sires and dams contributing to the cohort is broken down by age.

If, for instance, the maximum age of a dam in the whole data set is 14 years, the table will have 15 columns, one for each year class (and less than one year). Each column then contains the number of different sires and dams of the animals born in that year and age class. Furthermore, the average age of these parents is computed. This column gives a quick overview if the average age of dams and sires that have produced offspring has changed over the years.

Set 3: Distribution of parity

This table is closely related to the previous one, both in content and layout. With a length of the reproductive cycle of one year, the content will be very close to the previous one for females. With much shorter cycles, the parity-based statistics may be more useful.

Cohort: *The cohort is defined as the total number of animals born in a given year.* The total number of sires and dams contributing to the cohort is broken down by parity.

Set 4: Generation interval

The generation interval (GI) is one of the key factors affecting the rate of genetic progress per unit time. In the literature, the GI is computed in a number of ways with different levels of accuracy, yielding results that are not comparable. In its simplest form, GI is based on the actual or presumed average age of males and females in the herdbook. Here, we have decided to follow Falconer and Mackay's definition; they defined GI as the average age of the parents at the birth of their selected offspring (Falconer and Mackay, 1996). In the calculation of generation interval, an offspring is considered selected if it has produced at least one progeny. Also here, the GI is computed for each year.

Cohort: *The starting set for a cohort is all animals born in a given year (subset 1). Animals in subset 1 that become parents in later years are identified (subset 2). The parents of subset 2 are identified (subset 3). For each animal in subset 2, the average age of its parents at birth is computed. The GI is also computed for the four selection paths: sires to sons, sires to daughters, dams to sons, and dams to daughters.*

The generation interval is calculated separately for the males and females, along with the number of males and females that gave rise to the particular cohort for each of the four paths. The overall generation interval for the entire population is also provided.

Set 5: Variance of family size

Family size refers to the number of offspring of an individual that become breeding individuals in the next generation (Falconer and Mackay, 1996). The consequence of increased variation in family size is an increase in the rate of inbreeding and the reduction in the effective population size. Consequently, ensuring balanced usage of males and females in reproduction is a simple and efficient procedure to control inbreeding for a given population size (Groeneveld, 2003). The variance of family size can be minimized, i.e., regressed to zero, as the numbers of offspring become equal for all parents.

Cohort: *All animals born in the database are included in the cohort.*

The summary statistics for family size (i.e., the minimum, maximum and average) for the male and female parents are presented in the report. Offspring per animal are categorized into four groups as follows: all offspring born in the population and selected offspring for females and males separately.

In many breeding populations, individual animals are well known, with particularly popular ones having many offspring. This is presented in eight histograms, with the first block referring to all offspring and the second to selected offspring. The first histogram gives the number of offspring per animal ID sorted by number of offspring for the first 30 animals.

Large offspring groups are common in artificial insemination; but this statistic is also useful in embryo transfer programs, as it depicts the number of offspring generated by individual dams. The next set of histograms is based on selected offspring; it is more important in that it shows the actual contribution of an animal to future generations. This information can help balance usage of sires in the management of breeding populations.

Pedigree Analysis

Pedigrees are used extensively in animal breeding; the first herdbooks started as pedigree registers. Today pedigrees are the basis of BLUP genetic evaluation. Meaningful results can only be generated if they are correct. Areas addressed in the Pedigree Analysis Report are the quality of the pedigree, inbreeding-related statistics and those connected with additive genetic relationships. While the level of inbreeding is not of great importance, its rate of increase is the prime parameter in assessing the loss of additive genetic variation in a population. Here, the effective population size N_e is a major statistic. However, computation of N_e is anything but straightforward, with different assumptions leading to different estimates. This is why a number of approaches are implemented in the Pedigree Analysis Report.

Pedigree quality

The quality of pedigrees in production populations is not generally known. While herdbook societies have been founded on the bases of pedigree recording, considerable introgression has always occurred. This leads to a situation in which some animals have a much longer pedigree than others. As a result, in pedigree-based analyzes, animals with no known parents are assumed to be unrelated. Thus, as fewer ancestors are generally known, inbreeding will be underestimated.

A number of parameters can be used to assess what can be called "quality". Here, we chose the algorithm from MacCluer et al. (1983), which is a weighted completeness index.

This pedigree completeness index (PCI) summarizes the proportion of known ancestors in each ascending generation. It quantifies the chance of detecting inbreeding in the pedigree (Sørensen et al., 2005). The following formula was used to compute pedigree completeness (MacCluer et al., 1983):

$$I_d = \frac{4 I_{d_{pat}} I_{d_{mat}}}{I_{d_{pat}} + I_{d_{mat}}}$$

and

(Equation 1)

$$I_{d_k} = \frac{1}{d} \sum_{i=1}^{d} a_i \qquad k = pat, mat$$

where *k* represents the paternal (*pat*) or maternal line (*mat*) of an individual and a_i is the proportion of known ancestors in generation *i*. The *d* is the number of generations considered in the calculation of pedigree completeness. For example, if *d = 5*, then five ancestral generations

will be taken into account in the computations. The values for pedigree completeness range from 0 to 1. If all ancestors of an individual to some specified generation (*d*) are known, then $I_d = 1$, or if one of the parent (i.e., sire or dam) is unknown, $I_d = 0$.

Cohort: *A cohort is defined as all animals born in a given year.* The pedigree completeness values are presented as yearly averages.

Inbreeding

The relevant quantity for the assessment of inbreeding is the inbreeding coefficient *F* and its rate of change Δ*F*. The latter is defined by Falconer and Mackay (1996):

$$\triangle F = \frac{F_t - F_{t-1}}{1 - F_{t-1}} \qquad \text{(Equation 2)}$$

where *t* is the t*th* generation. As has been stated, the cohort is those animals born in a given year. To compute Δ*F*, two cohorts, i.e., sets of animals actually need to be defined: $Cohort_t$ at generation *t* and $Cohort_{t-1}$ at generation *t-1* have to be defined.

$Cohort_t$: *this cohort is defined as all animals recorded and born in a given year*.

$Cohort_{t-1}$ can be defined in two ways. The first definition is based on the actual parents of $Cohort_t$, while the second uses the average generation interval to arrive at a conceptual parents' birth year.

$Cohort_{1_{t-1}}$: *this cohort is defined as the set of parents of all animals in $Cohort_t$.* The average inbreeding coefficient of all animals in this cohort is computed to yield F_{t-1}.

$Cohort_{2_{t-1}}$: *this cohort is defined as all animals born one generation earlier than $Cohort_t$, based on the average generation interval as defined in Set 4.* The average inbreeding coefficient of all animals in this cohort is computed to yield F_{t-1}. This cohort will also include animals not related to the current $cohort_t$, whereas $Cohort_{1_{t-1}}$ only includes parents of $Cohort_t$.

Δ*F* is then computed for each year using formula 2. Because of the two ways to define $Cohort_{t-1}$, two estimates for Δ*F* will be obtained, hereon referred to as ΔF_p and ΔF_g for the estimate based on $Cohort_{1_{t-1}}$ and $Cohort_{2_{t-1}}$, respectively.

Alternatively, the rate of inbreeding can be calculated using log regression of (*1-F*) on birthdate (Pérez-Enciso, 1995). The average level of inbreeding in a population at a given time F_{u+t} relative to some time in the past (*u*) is given by the following formula:

$$(1 - F_{u+t}) = (1 - \Delta F)^t (1 - F_u) \qquad \text{(Equation 3)}$$

taking the natural log yields:

$$ln(1 - F_{u+t}) = t * ln(1 - \Delta F) + ln(1 - F_u) \qquad \text{(Equation 4)}$$

It is clear that when Δ*F* is constant, Equation 4 is a straight line with a slope *b* equal to -Δ*F*. Therefore, the rate of inbreeding per generation is:

$$\Delta F = (-1)bL \quad \text{(Equation 5)}$$

where L is the generation interval and b the slope from the regression of $ln(1\text{-}F)$ on year of birth.

Equation 4 is based on average inbreeding of the population at a given time. However, when individual inbreeding coefficients are available, as is the case with pedigree data, the rate of inbreeding can be calculated by regressing $ln(1\text{-}F_i)$ on year of birth, where F_i is the i*th* individual's inbreeding coefficient. The ΔF computed in this manner based on Equation 5 and using ΔF_i is from hereon referred to as ΔF_{ln}. Using F_i has the advantage that standard errors of the rate of inbreeding and effective population size could be calculated. Furthermore, individual inbreeding coefficients automatically accounts for the differences in the number of records per year.

Additive genetic relationship

The inbreeding coefficient and its rate of change depends on the mating decision of parents, reflecting an action in the previous generation. In contrast, the additive genetic relationship f (AGR) of a group of contemporary animals reflects the total average genetic relationship in that group. For instance, mating within herds will lead to a possibly large increase in the level of inbreeding, with a correspondingly high ΔF. The AGR across all herds, on the other hand, will still be low. The definition made above for the cohorts used to compute ΔF also applies here.

Cohort$_t$: *this cohort is defined as all animals born and recorded in a given year*. All conceptual "matings" of males and females within the birth year are made. The f is the overall mean of all conceptual "matings" of this cohort.

Cohort$_{t-1}$: *this cohort is defined as all animals born one generation earlier than Cohort$_t$ based on the average generation interval as defined in Set 4.* The f is the overall mean of all conceptual "matings" of this cohort.

For a set of potential parents a conceptual average genetic relationship can be computed, not only for the parent generation but also for the most recent, which may just be born.

Due to its conceptual nature, deciding on the actual "matings" implies a certain degree of arbitrariness.

We decided to make all possible "matings" among males and females of a cohort, i.e., a birth year. The balance of males and females in the cohort is dependent on the animal recording scheme or more precisely on the datafiles supplied. With complete registration of all animals born in a breeding program, as would be the case in most herdbook systems, each cohort would have roughly the same number of males and females. If, however, only the pedigrees of selected animals are supplied to POPREP, the cohorts will have much fewer males than females.

The development of AGR over generations can be expected to be more stable than the inbreeding coefficient because it is based on all possible matings, while the actual choice of matings can reduce the inbreeding coefficient of the offspring at the expense of having more closely related animals in the next generation.

The rate of change of the AGR Δf is analogous to Equation 2:

$$\triangle f_g = \frac{f_t - f_{t-1}}{1 - f_{t-1}} \quad \text{(Equation 6)}$$

Obviously, the computations for F and f are very different. For the latter, the cohort is split into the group of males and females. Then, the AGR is computed for every male "mated" to each female and averaged over the cohort. This procedure can be computationally intensive, as the computing time increases quadratically with the number of animals involved.

Effective population size

Different methods can be used to compute the effective population size. The following methods were implemented in POPREP.

Method 1: Assuming discrete generations and equal numbers of breeding males and females N_e can be determined directly as

$$N_e = \frac{4N_m N_f}{N_m + N_f} \qquad \text{(Equation 7)}$$

The number of males and females are usually easy to determine, therefore this expression is often used where other ways of computing N_e are not possible. However, the assumptions that the ratio between breeding males and females is 1:1 and that all individuals in a real population have an equal chance to contribute genetically to the next generation are usually not met. Therefore, Equation 7 tends to overestimate N_e considerably.

All other methods are based on the definition: N_e

$$\Delta F = 1/2N \qquad \text{(Equation 8)}$$

Substituting N_e for N will accommodate any breeding structure. Then, N_e can be computed as:

$$N_e = 1/2\Delta F \qquad \text{(Equation 9)}$$

In POPREP, the rate based N_e (Equation 9) is computed using different definitions of ΔF, as indicated above.

Method 2: Here, N_e is computed using ΔF_p. This means that for each year only the possibly rather small group of parents serves as a base for ΔF. It is to be expected that the resulting N_e will fluctuate, depending on the matings that resulted in the cohort of that birth year. On the other hand, with changing population sizes, this procedure will reflect the most recent estimate.

Method 3: Here, N_e is computed using ΔF_g. The ΔF_g is also based on animals that never contributed to the current generation, while it will also not include all parents of the current birth year. Compared to Method 2, more stable N_e estimates are expected.

Method 4: Here N_e is computed using ΔF_g. Under random mating Δf and ΔF are the same, a condition that tends not to be met in animal breeding. Avoidance of inbreeding and matings within herds will have a substantial impact on the rate of inbreeding, while this will

affect Δf to a much lesser degree. Further, since it is based on a large number of pseudo matings, the parameter is expected to be more stable. A final added advantage is expected because the Δf of the current generation reflects its population structure, while the ΔF reflects that of the previous generation.

Method 5: Here, N_e is computed using ΔF_{ln}. The advantage of this procedure is that it automatically yields the rate on the basis of individual inbreeding coefficients, assuming constant rate of change over the period of time considered. Methods 2 and 3 yield effective population sizes for each birth year. They implicitly also fit a regression, but only based on the two averages. If more years are to be considered together for an estimate of N_e, then this approach will give correct weights of the individuals involved.

Because ΔF_{ln} is based on the inbreeding coefficient of individuals, arbitrary time windows can be chosen for which N_e can be computed. An added advantage lies in the fact that standard errors in the rate are readily available.

Method 6: In Methods 2 through 4, it is assumed that the average inbreeding coefficient for generation t can be readily computed, i.e., that discrete generations can be defined. That this is not straightforward is indicated by the need for two procedures (Methods 2 and 3) to define the cohort of a generation. In the case of overlapping generations, the following formulae can be used to compute N_e (Pérez-Enciso, 1995):

$$(1 - F_t) = (1 - 1/2N_e)^t \qquad \text{(Equation 10)}$$

therefore:

$$N_e = 1/2(1 - (1 - F_t)^{(1/t)}) \qquad \text{(Equation 11)}$$

where F_t is the average inbreeding coefficient for a specific cohort (cohort is defined as a group of animals born within the same year) and t the number of generations it took for the population to reach F_t. The year prior to the birth year of the first cohort for which an generation interval could be calculated (see definition for Set 2) was assumed to be the base year. Based on the generation interval for each cohort and the distance between the current cohort and the base year, t was estimated.

POPREP Customization and Outputs

The population reports are generated as pdf documents, ready to be printed. They contain text blocks, tables and graphs. Furthermore, intermediate results are also available for further investigations in the form of comma-separated values (csv) files.

Customization

In the released software, the text blocks in the reports describe definition, computation and meaning of the various parameters in English. These texts are stored in ASCII files, and thus they can be easily changed, depending on the intended use. The default version can be used out of the box, if the intention is to compute the parameters for a certain population. On

the other hand, it may make sense to provide breed-specific reports at regular intervals, for instance as a part of a breed society's yearly reporting, as suggested by Groeneveld (2003). Here, the default texts can be replaced by whatever is deemed appropriate for a particular breed. Such breed-specific text can then be used with little modification, and with nearly automatic production for the yearly breed reports.

Outputs

While the default output from this system is two automatically generated typeset reports in pdf format, the system also provides the actual numerical data that were used to generate them. For each table, an ASCII file is provided with csv, which can be loaded, for instance, into spread sheets. Apart from these 22 csv files, the inbreeding coefficent for each animal is supplied together with the year of birth. Along with the program *log_of_inbreeding.pm*, the users can define an own-time range for the computation of effective population sizes, together with the specification of the pedigree completeness index for a given pedigree depth. Furthermore, the $ln(1\text{-}F)$ inbreeding coefficients of each animal with its PCI are dumped to a file. In general, provision of intermediate data allows users to conduct further analyses that may go well beyond the outputs generated by the reports.

Implementation, Scaling and Availability

In the true spirit of Open Source (Raymond, 2001), the package is written relying heavily on other freely available software. It is integrated in the APIIS framework (Groeneveld, 2004). This, in turn runs on Linux with Postgresql (Momjian, 2001) as the database and Perl as the major programming language (Wall and Schwartz, 1991). Typesetting is done with LaTeX. Graphics are created through gnuplot.

Both reports are designed as stand-alone batch applications that run with an APIIS database, picking up those breeds that are stored in the database. No preprocessing or data selection is required, i.e., the report generation is completely self-contained and automatic.

The procedure has two parts. In the first step, all numerical values are generated and, where required, stored in the database. This is done in Perl, using standard SQL to query the database. In the second step, the final reports are typeset using LaTeX for high quality output. This step merges predefined explanatory text, outlining the problem and giving the computational procedures and formulae for each of the statistics and the final data. Furthermore, all data are also exported to comma-separated files, so that they can be processed at liberty after the report has been generated by whatever software the user chooses.

The inbreeding coefficients are computed for all animals in the database, using the fast tabular method and then stored temporarily in the database for further computations. The additive genetic relationships are computed using the PEDIG Fortran package (Boichard, 2002) and specifically the *par3.f* program, which we modified to fit in the workflow.

Computational issues

The most CPU-intensive tasks are computation of the inbreeding coefficient (F),

PCI and additive genetic relationship (*f*). Computing the additive genetic relationship of all animals in a population does become a computational burden, as each cohort, i.e., birth year together with its own pedigree, is processed separately. The same applies to the PCI. Precisely this independence of cohorts allows parallelization of the complete process, here implemented in Perl. Depending on the number of processors available on the computer (or the choice made by the user) the program will open an equal number of independent threads. The scaling of the algorithm is nearly perfect because of the coarse granularity of our problem: depending on the size of the population computing the average genetic relationship for a cohort may take minutes and much more before control is given back to the scheduler.

Availability

POPREP is released under GNU Public License (GPL) Free Software Foundation, Inc. (1991). There are three ways to install/use POPREP. Firstly, the POPREP software comes with every APIIS installation, and is thus, after installation of the complete system, available on any permanent APIIS database. Secondly, the reports can also be generated by loading ASCII pedigree data temporarily for the purpose of the report into an APIIS-conforming database, which the user installs on his own hardware through an appliance, which will then be run as a virtual machine on the user's own computer. A third alternative is the website (http://poprep.tzv.fal.de), which is provided as a service to the animal-breeding community. After successful uploading of the data by the user, checking of the data is started on the web server, followed by the computation of the population report. In case of input errors as well as on successful completion, the user is notified through e-mail and will receive either the data error list or the completed population report.

Investigation of Eight Breeds

To investigate the potential of POPREP, it was run on a number of populations from different species. It was the objective to select populations covering a wide range of breeding programs involving pigs, dairy and beef cattle, and sheep, ranging from intensive commercial situations to an endangered breed. The populations were the beef breeds Nellore from Brazil and the Bonsmara from South Africa, dairy cattle breeds Jersey and Holstein from South Africa, the pig breeds Duroc and Landrace, also from South Africa, and finally the sheep breeds Merino and the endangered Skudden from South Africa and Germany, respectively.

The outputs presented are just a subset of those available from POPREP. They are given as examples, with hints on how the outputs can be used further.

An overview of the populations is given in Table 1. The population sizes varied considerably from 12,400 for the Skudden to more than two million for the Holstein dairy population. The column "period" gives information about the data structure. The beginning of substantial pedigree recording in the datasets also varies widely from 1950 for the Holsteins to 1984 for the Nellore. The last year of complete performance records varied from 2007 to 2008.

Table 1. General statistics.

Breed	Period 1st/begin recording	Total number of animals	Animals in reproduction	
			Males	Females
Nellore	1960/1984-2007	483,291	1,663	161,440
Bonsmara	1954/1970-2007	1,291,165	15,673	319,26
Jersey	1942/1975-2008	806,923	9,062	310,508
Holstein	1932/1950-2008	2,316,559	14,393	838,492
Duroc	1982/1982-2008	123,990	1,740	5,313
Landrace	1976/1982-2008	328,343	3,971	12,664
Merino	1980/1980-2008	197,463	3,046	66,552
Skudden	1982/1990-2008	12,400	722	2,974

Landrace and Duroc

South African Landrace pigs were first imported into the country in 1952, while the first imports for the South African Duroc pigs took place in 1980. The SA Landrace originated from Holland, Sweden and Germany and Duroc from Canada. All South African pig breeds are registered at the South African Studbook association and there is litter and performance recording at the Agricultural Research Council's Animal Production Institute. In South Africa, all piglets are registered at birth as animals in the national database. This results in a roughly equal male and female population in the database. Frozen semen is imported on a yearly basis for both breeds. All importations, both frozen semen and live animals, require a three-generation pedigree. Accordingly, both pig breeds have a good pedigree completeness up to three generations, but for four generations and more, the completeness of the pedigrees declines as a result of regular importations.

Nellore

The Nellore population belongs to a company that started a breeding program in the 1980's and that today leads bull sales in Brazil, selling more than 2000 bulls/year. This particular population was composed of animals not registered by the official breed association. In the 1990s, the program got the approval of the Brazilian Ministry of Agriculture, and could start sell animals with a special document called "CEIP - Certificado Especial de Identificação e Produção", which allows superior breeding stock from commercial herds to attain the same status as the seedstock sector. The program has today close to 18,000 cows and uses both artificial insemination and natural breeding. Animals are reared in tropical pasture conditions in farms located in savanna-type regions called "cerrado". Bulls are sold at an age of two years and cows are bred when they are 14 months old, or during the next breeding season, when they are close to 2 years old. Commercial animals are reared in pastures and slaughtered between 24 and 36 months of age. Only a small proportion of animals are finished in feedlots, for 90 days, with medium-energy, corn-silage-based diets.

The base population came from commercial animals, with no pedigree information. As the company had a demand for more than 500 replacement bulls per year and could not find genetically evaluated bulls, a selection program was established in the early 1980's. Due to large farms and pasture pen sizes, groups of large numbers of cows are still being bred by multi-sire groups of bulls. That has major implications on pedigree completeness. Also, there

are bulls used to a great extent in the population that has no pedigree information in early generations, a situation that is even more common with cows.

Holstein, Jersey, Bonsmara, and Merino

Animal registration in South Africa is the responsibility of the South African Studbook and in some instances that of the breed societies. All animal registration information must be captured in the national database - Integrated Registration and Genetic Information System (INTERGIS). The Holstein, Jersey, Bonsmara, and Merino breeds are managed in open herdbooks. This implies that animals may enter the herdbook without pedigrees. In the case of imports, a three-generation pedigree is required.

Skudden

Skudden sheep are an old endangered German sheep breed. It can thus be taken as a model for the non-mainstream breeds, which are kept as genetic resources in a conservation breeding program. Often, animals are bred by dedicated individuals, who keep the population alive. While data collection and management typically have great importance in commercial breeding programs, these do not necessarily have to be so for endangered breeds and the operational environment that they are kept in. The effects of this may be less complete pedigrees.

RESULTS FOR ACTUAL POPULATIONS

The results presented here are a subset of the results returned by POPREP. The emphasis is placed on the generation interval, pedigree completeness, rate of inbreeding, and effective population sizes. For a complete sample report, see http://poprep.tzv.fal.de. The post-processing options were investigated to indicate possibilities as well as areas a user should be concerned about.

Generation Intervals

The exact generation intervals are listed in Table 2. These are from the last line of the table on generation intervals of the Population Structure Report.

Table 2. Generation interval (GI) in years.

Breed	Paths				Average
	s-s	s-d	d-s	d-d	GI
Nellore	10.1	7.2	8.4	5.9	6.1
Bonsmara	5.4	5.5	5.7	5.7	5.6
Jersey	7.9	6.6	5.7	4.7	5.7
Holstein	8.2	8.0	5.1	4.7	6.5
Duroc	1.9	1.8	1.8	1.7	1.8
Landrace	2.0	1.9	2.1	2.0	2.0
Merino	2.7	2.6	3.2	3.0	2.8
Skudden	4.1	4.0	3.0	3.7	3.8

s-s = sire-son; s-d = sire-daughter; d-s = dam-son; d-d = dam-daughter.

In the report, the generation interval is not only computed for the current generation but independently for each birth year, which gives the user the possibility to observe the changes in the various paths over time. Additionally, the report gives the number of selected animals in each path for each year, giving an idea of the precision of that estimate. Through this information, goals in a breeding program can be checked against reality.

Pedigree Completeness

Table 3 gives the pedigree completeness for the populations. Pedigrees are the basis for BLUP genetic evaluation in animal breeding. Likewise, pedigrees are the basis for the assessment of the inbreeding structure in populations. However, often pedigrees are not defined in their quality, and the degree of incompleteness for the parameter to be estimated (BLUP or inbreeding) is unknown. As can be seen in Table 3, the PCI shows big differences among the populations. The figures reported are from the last complete birth year as given in the report. Even the rather low requirement of three generations backwards yields a complete pedigree only for the Duroc and Landrace, with only about 70% completeness for Jersey, Holstein and Merino. The very low value of about 29% for Nellore indicates a continuing heavy influx of new animals with apparently little pedigree data. With this information, the user can investigate if the pedigrees are good enough for their intended purpose. Clearly, in the case of the Nellore, inbreeding will be underestimated if immigration into the recording scheme originates from the general pool of animals that will likely not be unrelated. What the actual status is, only further investigations about the population can reveal; this cannot be answered by the report. Its objective is restricted to pointing to a potential problem. As the completeness index is computed for each birth year for pedigrees depths of 1 through 6, the user can follow the development, and hopefully the improvement made in recording.

Table 3. Pedigree completeness index (PCI) for the last birth year.

Breed	Pedigree completeness index				
	PCI2	PCI3	PCI4	PCI5	PCI6
Nellore	0.31	0.29	0.26	0.23	0.20
Bonsmara	0.87	0.84	0.82	0.80	0.78
Jersey	0.76	0.74	0.74	0.70	0.67
Holstein	0.75	0.71	0.65	0.59	0.53
Duroc	1.0	0.99	0.98	0.96	0.94
Landrace	1.0	1.0	0.98	0.97	0.95
Merino	0.89	0.72	0.59	0.49	0.41
Skudden	0.67	0.52	0.42	0.36	0.31

PCI 2-6 = PCI for pedigree depths of 2 to 6 generations.

Rates of Inbreeding

The rates of inbreeding per generation calculated using different equations are given in Table 4. Generally, the rates of inbreeding based on inbreeding for the two definitions of $cohort_{t-1}$ are not consistent. For example, the rates of inbreeding within breed had different signs for the Holstein and Duroc. However, it should be noted that the rates of inbreeding were somewhat similar for Landrace.

Table 4. Average rate of inbreeding and additive genetic relationship for the last generation.

Breed	ΔF_p	ΔF_g	Δf_g	ΔF_{ln}
Nellore	0.0000	0.0000	-0.0001	-0.0002
Bonsmara	0.0033	0.0001	0.0014	0.0011
Jersey	0.0007	0.0103	0.0020	0.0099
Holstein	-0.0016	0.0005	0.0013	0.0000
Duroc	0.0089	-0.0058	0.0012	-0.0007
Landrace	0.0189	0.0169	0.0074	0.0250
Merino	0.0067	0.0031	0.0008	0.0038
Skudden	-0.0193	-0.0072	-0.0010	-0.0011

Effective Population Size

Based on the output from POPREP, estimates for all the methods provided were obtained on for the years 1990-2008, with the exception of the Bonsmara and Nellore breeds, for which data were only available until 2007, as shown in Table 5. It should be noted that these N_e are not based on the rates of inbreeding in Table 4. For Methods 1 through 4, the yearly values as computed were averaged. For Method 5, the ΔF was computed for all animals on the basis of Equation 5, with the restrictions being a pedigree completeness of 0.8 for a five-generation pedigree.

Table 5. Effective population sizes based on the period 1990-2008.

Breed[1]	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6[2]
Nellore	502	-[3]	-	-	476	1293
Bonsmara	6901	106	162	324	190	231
Jersey	3545	203	79	83	67	208
Holstein	4803	-	318	282	408	794
Duroc	247	40	119	99	219	143
Landrace	546	50	13	44	222	102
Merino	861	241	562	5588	447	601
Skudden	280	-	528	391	-	210

[1]Five generations with minimum pedigree completeness index 0.8; based on arithmetic mean of ΔF_p, ΔF_g and Δf_g. [2]N_e was taken from the most recent full birth year. [3]Negative.

As expected, estimates among populations were substantial. However, the differences among the methods were even larger. Differences among parameter estimates are to be expected, because they estimate the N_e only under certain and different assumptions, as can be seen from Equations 2 and 6. Assessing the methods for their applicability to estimate the effective population sizes, it is only clear that Method 1 is not appropriate for breeding populations under selection. The remaining methods also show considerable variation. For the Nellore, with their low degree of pedigree information, three methods even produce negative estimates. The N_e for the Merinos shows an even wider range, from 241 to 5511. The estimates for the Bonsmara range between 106 and 324 for Methods 2 to 6.

While the years to be covered can be freely chosen for Methods 1 through 5, Method

6 is dependent on a predefined starting point or base year. Because of the fact that ΔF is tied to a "base" or foundation point of the population, with an inbreeding coefficient of zero, Figure 1 shows that the N_e obtained by using this formula are more stable from one year to another. Due to the rate being estimated from the base or zero point, a negative rate of inbreeding is not possible. Therefore, an N_e can always been estimated if the average inbreeding coefficients of the current cohort are greater than zero.

However, as is the case with the other methods, the starting year also has to be defined. This is critical, because the N_e calculations depend on the interval between the base year and the birth year of the cohort, which is the basis for calculating the average inbreeding coefficient. Although this method presents more stable results, it is also dependent on the quality and completeness of the herdbook. Figure 1 clearly indicates the effect of pedigree quality. The N_e estimates for the Duroc, Landrace, Holstein, Jersey, and Bonsmara breeds are relatively stable; unfortunately for the breeds with less complete pedigrees (Merino and Nellore), N_e estimates varied much more.

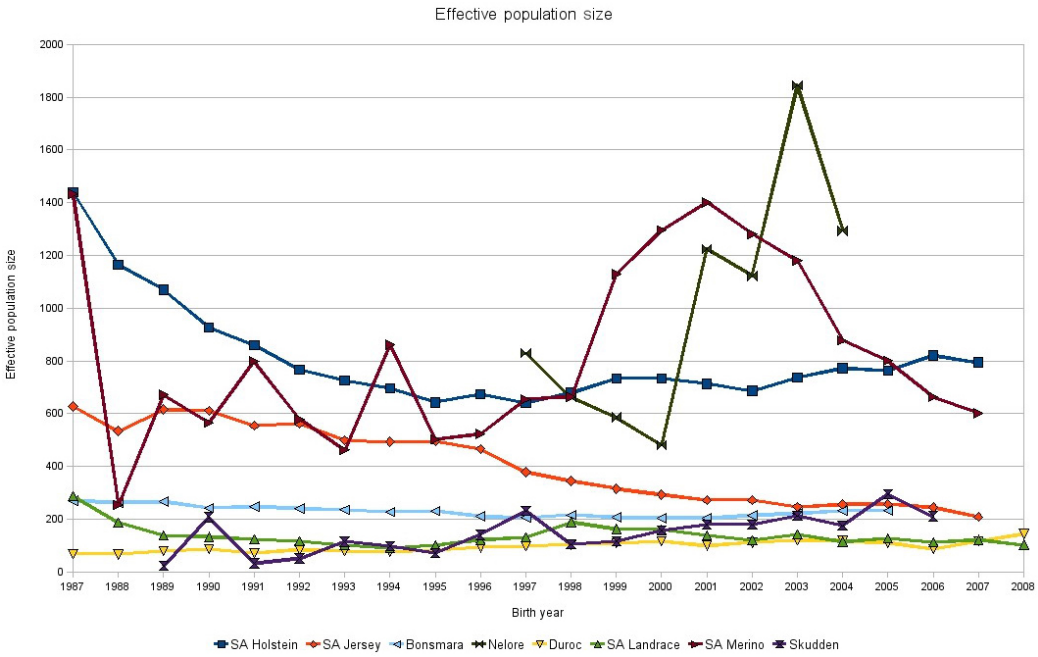

Figure 1. Trend of N_e computed using Method 6.

As seen, even if a method is chosen, further decisions have to be made to obtain a final estimate. This is shown by the data in Figure 2 for the Landrace and Bonsmara dataset. Here, N_e was estimated using Method 5 and computed according to formula 5 on the basis of pedigrees for four, five and six generations deep restriction, requiring 0, 40, 60, 80, 90, 95% pedigree completeness. For Landrace, the effective population size seems to increase with pedigree completeness, in the case of a six-generation deep pedigree going from 187 to nearly

450. Obviously, the number of animals in the analysis goes down as the restrictions increase; for Landrace this is still 80,000.

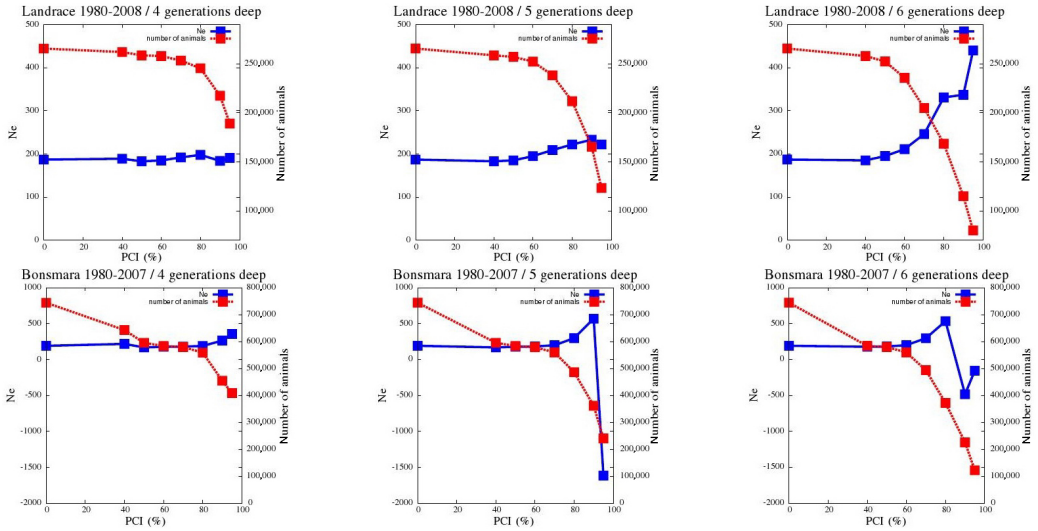

Figure 2. Landrace and Bonsmara pedigree completeness index (PCI) and N_e for four, five and six generations pedigree depth.

The Bonsmara breed in Figure 2 shows a slightly different pattern; initially, the N_e increase with PCI on the basis of 4 generations. However, for five and six generations, the increase is only up to 0.8 for five generations and 0.9 for six generations, whereafter it drops to a negative number. The reasons for this may be related to differential selection over the years.

At this stage, the reasons for the change in N_e as a function of pedigree completeness can only be speculated about. But what has become clear, is that a number of relatively arbitrary decisions have to be taken to compute an N_e with any method. These decisions will greatly influence the estimate, but are seldom reported. Cervantes et al. (2008) investigated a number of procedures for estimating N_e and also found a large variation in the estimates for real pedigrees using the ENDOG tool (Gutierrez and Goyache, 2005). Clearly, more investigations are required; here POPREP with its post-processing option can be a useful tool.

Computational Requirements

Depending on the size and depth of the pedigree, the computational requirements can be substantial. Table 6 gives the run time for the most computationally intensive blocks and the total wall clock time it took for completion of the reports for the breeds investigated.

Table 6. Run time of the population report in minutes.

Breed	Total number	*F*	PCI[1]	*f*	Total
Nellore	483,291	02:19	02:44	10:58	0:18:58
Bonsmara	1,291,165	12:19	15:03	260:34	5:03:25
Jersey	806,923	22:22	08:37	30:47	1:10:19
Holstein	2,316,559	20:54	81:18	31:16	2:36:21
Duroc	123,990	01:07	11:01	08:26	0:22:06
Landrace	328,343	03:02	07:26	47:17	1:01:15
Merino	197,463	01:42	00:29	03:23	0:07:31
Skudden	12,400	00:04	00:02	00:01	0:00:26

[1]Pedigree completeness index and *f* run parallel on 16 (Nellore and Skudden 8) processors.

With population sizes in the millions, hours of computation may be expected. It must be noted that the most computationally demanding parts were run on multiprocessor computers with eight or 16 processors. On a single processor machine, the total time to completion will be substantially longer.

DISCUSSION

Pedigree data on several livestock breeds from different species were used to demonstrate the potential of POPREP. Results from the current investigation showed that the time required to generate the reports could be minimal for small pedigrees (e.g., <1 min and 10 min for the Skudden and Merinos, respectively). Computing time can be substantial for large pedigrees. For the purpose of monitoring breed diversity, the reports from POPREP may be required periodically, e.g., once a year. Therefore, computing time might not be a major issue where reports from POPREP are required for monitoring purposes. While pedigrees larger than those considered in the current study are a possibility in practice, the computing time should in general not be an issue.

POPREP generates two reports, i.e., the Population Structure and Pedigree Analysis Reports. Results in the Population Report are straightforward and therefore easy to use for population monitoring. This is due to the fact that the parameters presented in the Population Structure Report could be easily computed from pedigree data and birth dates. The results are presented on a yearly basis to allow for assessment of the evolution of the parameters, which is more important for monitoring purposes. On the other hand, several estimates of the same parameter (e.g., rate of inbreeding and effective population size) are provided in the Pedigree Analysis Report. Different estimates of the same parameter result from the fact that different equations invoke different assumptions. In open populations, for example, negative effective population size is possible as a result of immigrants into the population.

In POPREP, different approaches of computing the parameters were implemented to allow for comparison of results. Estimates of the effective population size from different methods within breed are not consistent in general (see Table 5). These results point to the dilemma facing population managers in terms of which estimate (e.g., effective population size) to consider since different estimates may call for different conservation actions.

POPREP allows the user to set the number of generations and pedigree completeness for which the rate of inbreeding and effective population size should be computed. Figure 2

shows results where such restrictions were used. Of interest is that the effective population size depends on the number of generations set. For both the Landrace and Bonsmara breeds, the effective size showed an erratic behavior at pedigree completeness greater than 80%. For the Bonsmara, the effective population size ranged from a positive value at lower pedigree completeness to a negative effective population size at pedigree completeness greater than 80%. The results from other breeds (not shown) also showed a similar pattern. The possible explanation for such results is unknown. It is not known, however, which effective population size should be considered.

Methods 5 and 6 are closely related, as they are based on the same expression (Equation 10). The former yields on estimate for a given starting and end point, assuming a constant N_e over the period, while the latter yields one N_e estimate for each birth year, as shown in Figure 2. The average F of each birth year also contains the accumulated inbreeding of the ancestors, but still the N_e estimates change with time, as can be seen by the Holstein and Jersey curves in Figure 2. In contrast, Method 5 will produce only one estimate for the whole period, unless it is repeatedly run for each birth year using the same starting point. Whether Method 5 can be used for monitoring changes in population size is unclear and needs further attention.

In general, use of N_e for monitoring breeding populations is anything but straightforward, as it needs to record the changes in N_e in the last years, while estimation of N_e over a longer period is already wrought with complications.

CONCLUSIONS

Management of animal genetic resources in order to minimize loss of genetic diversity both within and across breeds has recently received attention at different levels, e.g., breed, national and international levels. Among the major impediments to successful conservation of genetic diversity is availability of timely and accurate information on the status of populations with respect to key parameters, such as generation interval, rate of inbreeding and effective population size. POPREP provides comprehensive reports on estimates of population parameters that could be used by decision makers, such as breed associations, conservation groups or government agencies for monitoring purposes. The reports generated by POPREP could be used in the annual evaluation of conservation strategies of individual breeds or prioritization of conservation actions by government. POPREP has the added functionality of providing output files in the form of cvs files that could be used for further processing.

ACKNOWLEDGMENTS

Financial support from the Federal Ministry of Food, Agriculture and Consumer Protection, Germany (BMELV), the Agricultural Research Council of South Africa (ARC) and the German Academic Exchange Service (DAAD) is gratefully acknowledged. We thank the Zuchtverband für Ostpreußische Skudden und Rauhwollige Pommersche Landschafe e.V. for sharing their data. Without the contribution of the following persons, POPREP could not have been developed: Helmut Lichtenberg, Lina Yordanova, Ulf Müller, Ralf Fischer, and Zhivko Duchev. Japie v.d. Westhuizen and Zhivko Duchev are thanked for their critical review of the draft manuscript.

REFERENCES

BMVEL - Bonn (2004). Tiergenetische Ressourcen: Nationales Fachprogramm/Hrsg: Bundesministerium für Verbraucherschutz, Ernährung und Landwirtschaft. Available at [http://www.bmelv.de/cae/servlet/contentblob/376770/publicationFile/22036/TiergenetischeRessourcen.pdf]. Accessed September 15, 2009.

Boichard D (2002). PEDIG: a Fortran Package for Pedigree Analysis Suited for Large Populations. In: Proceedings of the 7th World Congress on Genetics Applied to Livestock Production (WCGALP). Available at CD-ROM communication No. 28-13, Montpellier.

Cervantes I, Goyache F, Molina A, Valera M, et al. (2008). Application of individual increase in inbreeding to estimate realized effective sizes from real pedigrees. *J. Anim. Breed. Genet.* 125: 301-310.

Falconer DS and Mackay TFC (1996). Introduction to Quantitative Genetics. 4th edn. Longman, Essex.

Free Software Foundation Inc. (1991). GNU General Public License, Version 2, Copyright (C) 1989, 1991. 59 Temple Place - Suite 330. Free Software Foundation, Boston. Available at [http://www.gnu.org/copyleft/gpl.html]. Accessed September 15, 2009.

Groeneveld E (2003). Strategie und Logistik zur verantwortungsvollen Verwaltung der genetischen Diversität in der Nutztierzüchtun. *Züchtungskunde* 75: 317-323. ISSN 0044-5401.

Groeneveld E (2004). An adaptable platform independent information system in animal production: framework and generic database structure. *Livest. Prod. Sci.* 87: 1-12.

Gutierrez JP and Goyache F (2005). A note on ENDOG: a computer program for analysing pedigree information. *J. Anim. Breed. Genet.* 122: 172-176.

MacCluer JW, Boyce AJ, Dyke B, Weitkamp LR, et al. (1983). Inbreeding and pedigree structure in Standardbred horses. *J. Hered.* 74: 394-399.

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

Momjian B (2001). PostgreSQL: Introduction and Concepts. Addison-Wesley, Reading.

Pérez-Enciso M (1995). Use of the uncertain relationship matrix to compute effective population size. *J. Anim. Breed. Genet.* 112: 327-332.

Raymond ES (2001). The Cathedral & the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary. O'Reilly Media Inc., Beijing. ISBN 0596001088.

Sorensen AC, Sorensen MK and Berg P (2005). Inbreeding in Danish dairy cattle breeds. *J. Dairy Sci.* 88: 1865-1872.

United Nations (1992). United Nations Environment Programme Convention on Biological Diversity. Available at [http://www.cbd.int]. Accessed September 15, 2009.

Wall L and Schwartz RL (1991). Programming Perl. O'Reilly & Associates, Sebastopol.

