ORIGINAL ARTICLE

A note on ENDOG: a computer program for analysing pedigree information

J.P. Gutiérrez[1] & F. Goyache[2]

1 Departamento de Producción Animal, Facultad de Veterinaria, Avda. Puerta de Hierro s/n, Madrid, Spain
2 SERIDA-Somió, C/Camino de los Claveles, Gijón (Asturias), Spain

Correspondence
Juan Pablo Gutiérrrez, Departamento de Producción Animal, Facultad de Veterinaria, Avda. Puerta de Hierro s/n, E-28040-Madrid, Spain. Tel/Fax: +34 913943767;
E-mail: gutgar@vet.ucm.es

Received: 11 August 2004;
accepted: 22 November 2004

Summary

The aim of this note is to describe the program ENDOG (v.3.0). The program handles pedigree information to conduct several demographic and genetic analyses including: (a) the individual inbreeding and average relatedness coefficients; (b) effective population size; (c) parameters characterizing the concentration of both gene and individuals origin such as the effective number of founders and ancestors, the effective number of founder herds; (d) *F* statistics and paired genetic distances for each subpopulation under study; (e) descriptors of the genetic importance of the herds in a population and (f) generation intervals. The program will help breeders and researchers to monitor the changes in genetic variability and population structure with limited costs of preparing datasets. The program, user's guide and example file can be downloaded free of charge from the World Wide Web at http://www.ucm.es/info/prodanim/Endog30.zip.

Introduction

The assessment of the within populations genetic variability has received increasing attention over recent years (Woolliams *et al.* 2002). Considering both selection and conservation, some simple demographic parameters have a large impact on the evolution of the genetic variability and largely depend on the management of the population (Goyache *et al.* 2003; Gutiérrez *et al.* 2003; Honda *et al.* 2004). Moreover, breeders and researchers can be interested in the ascertainment of the extent in which an inappropriate mating policy leads to structuring the populations under study (Caballero & Toro 2002). Some computer routines are available to test the evolution of the genetic variability of populations using pedigree information (Boichard 2002). However little efforts have been devoted to pedigree analysis software. ENDOG (current version 3.0) is a population genetics computer program that conducts several demographic and genetic analyses on pedigree information in a friendly user's environment. ENDOG is tributary of a suite of FORTRAN 77 routines which were widely distributed and used among Spanish groups (Gutiérrez *et al.* 2003). ENDOG has been written in VisualBasic^{TM} language and runs under Windows 95/98/2000/NT/XP versions. A setup menu will guide users when installing the program. The program, user's guide and example file can be downloaded free of charge from the World Wide Web at http://www.ucm.es/info/prodanim/Endog30.zip.

Methods

Primary functions carried out by ENDOG are the computation of the individual inbreeding (*F*) (Wright 1931) and the average relatedness (*AR*) (Goyache et al. 2003; Gutiérrez *et al.* 2003) coefficients.

F is defined as the probability that an individual has two identical alleles by descent, and is computed following Meuwissen & Luo (1992). The increase in inbreeding (ΔF) is calculated for each generation by means of the classical formula $\Delta F = (F_t - F_{t-1})/(1 - F_{t-1})$, where F_i is the average inbreeding at the ith generation. Using ΔF, ENDOG computes the effective population size (N_e) as $N_e = 1/(2\Delta F)$ for each generation having $F_t > F_{t-1}$ to roughly characterize the effect of the remote and close inbreeding. N_e is defined as the number of breeding animals that would lead to the actual increase in inbreeding if they contributed equally to the next generation. Whatever the way to compute N_e, this parameter fits poorly to real populations in small populations with shallow pedigrees, giving an overestimate of the actual effective population size (Goyache *et al.* 2003). To better characterize this, ENDOG gives three additional values of N_e by computing the regression coefficient (b) of the individual inbreeding coefficient over: (i) the number of full traced generations; (ii) the maximum number of generations traced and (iii) the equivalent complete generations (Maignel *et al.* 1996), and considering the corresponding regression coefficient as the increase in inbreeding between two generations ($F_t - F_{t-1} = b$), and consequently (assuming $1 - F_{t-1} \approx 1$) $N_e = 1/2b$. When the available information is scarce, these estimations can be useful to approximate the upper (using i), lower (ii) and 'real' (using iii) limits of N_e in the analysed population.

The average relatedness coefficient (*AR*) of each individual is defined as the probability that an allele randomly chosen from the whole population in the pedigree belongs to a given animal. *AR* can then be interpreted as the representation of the animal in the whole pedigree. The description of the algorithm used to compute *AR* is given in Table 1. As shown in this Table it is possible to obtain the *AR* coefficients at the same time as the *F* coefficients by only writing an additional code line without increasing substantially the computational costs. Colleau (2002) recently presented an algorithm useful, among other things, to obtain the average relationship coefficients between each member of a group and the whole group (including self-relationships) and the average pairwise relationship coefficients. The algorithm implemented in ENDOG is equivalent to that of Colleau (2002) when the whole population is considered as a single group.

The advantages of using *AR* are: (a) the computational cost to calculate *AR* coefficients is similar to that for the computation of the numerator relationship matrix, because both procedures use common algorithms; (b) the *AR* of a founder indicates the percentage in which this founder can be consider the origin of the population; (c) *AR* coefficients can also be used as a measure of inbreeding of the whole population, as it takes into account both the inbreeding and the coancestry coefficients; (d) *AR* can be used as an index to maintain the initial genetic stock selecting as breeding animals those with the lowest *AR* value and (e) *AR*, as an alternative or complement to *F*, can be used to predict the long-term inbreeding of a population because it takes into account the percentage of the complete pedigree originated from a given founder at population level. In addition, *AR* can be used to compute the effective size of the founder population as the inverse of the sum of the square *AR* coefficients across founder animals.

Table 1 Description of the algorithm used in ENDOG to compute the individual average relatedness (*AR*) coefficients

Let a vector c' defined as:

$$c' = (1/n)1'A \quad (1)$$

A being the numerator relationship matrix of size $n \times n$. On the other hand, the numerator relationship matrix can be obtained from the P matrix, where $p_{ij} = 1$ if j is parent of i, and 0 otherwise, which sets the parents of the animals (Quaas 1976), by:

$$A = (I - \frac{1}{2}P)^{-1}D(I - \frac{1}{2}P')^{-1} \quad (2)$$

where D is a diagonal matrix with non-zero elements obtained by:

$$d_{ii} = 1 - \frac{1}{4}a_{jj} - \frac{1}{4}a_{kk} \quad (3)$$

j and k being the parents of the individual i. From 2,

$$A(I - \frac{1}{2}P') = (I - \frac{1}{2}P)^{-1}D$$

Premultiplying by $(1/n)$ $1'$:

$$(1/n)1'A(I - \frac{1}{2}P') = (1/n)1'(I - \frac{1}{2}P)^{-1}D$$

and using 1:

$$c'(I - \frac{1}{2}P') = (1/n)1'(I - \frac{1}{2}P)^{-1}D$$

Multiplying c' into the parenthesis and isolating c':

$$c' = (1/n)1'(I - \frac{1}{2}P)^{-1}D + \frac{1}{2}c'P' \quad (4)$$

As the computation of both A and the *AR* coefficients involves the term $(I - \frac{1}{2}P)^{-1}$ D, it is possible to obtain the *AR* coefficients at the same time as the *F* coefficients by only writing an additional code line without increasing substantially the computational costs.

At the moment of the computation of *F* and *AR* coefficients, ENDOG computes for each individual the number of full generations traced, the maximum number of generations traced and the equivalent complete generations for each animal in the pedigree data. The first is defined as the furthest generation in which all the ancestors are known. Ancestors with no known parent were considered as founders (generation 0). The second is the number of generations separating the individual from its furthest ancestor. The equivalent complete generations is computed as the sum over all known ancestors of the terms computed as the sum of $(1/2)^n$ where n is the number of generations separating the individual to each known ancestor (Maignel *et al.* 1996).

Using ENDOG it is possible to assess the concentration of the origin of both animals and genes computing the following parameters: (a) effective number of founders (f_e); (b) effective number of ancestors (f_a) (Boichard *et al.* 1997) and (c) effective number of founder herds (f_h). The first is defined as the number of equally contributing founders that would be expected to produce the same genetic diversity as in the population under study. This is computed as:

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

where q_k is the AR coefficient of the founder k. Parameter f_e, as computed by ENDOG, would be equivalent to that computed following Lacy (1989) if the reference population used is the whole pedigree. Parameter f_a is the minimum number of ancestors, not necessarily founders, explaining the complete genetic diversity of a population. The parameter f_a complements the information offered by the effective number of founders accounting for the losses of genetic variability produced by the unbalanced use of reproductive individuals producing bottlenecks. It is computed in a similar way to the effective number of founders:

$$f_a = 1 \Big/ \sum_{j=1}^{a} q_j^2$$

where q_j is the marginal contribution of an ancestor j; in other words, the genetic contribution made by an ancestor that is not explained by other ancestors chosen before. The last two parameters are initially computed by ENDOG using as reference population all the individuals in the pedigree with both parents known. However they can be recomputed after choosing a particular reference population. The effective number of herds is simply computed as the inverse of the summed squared of the sum of the contributions of the Boichard *et al.*'s (1997) ancestors into each herd.

ENDOG can be used to infer population structure from pedigree information. ENDOG can compute Nei's minimum distance (Nei 1987) and F statistics (Wright 1978) for each predefined subpopulation (i.e. according to sex, areas, herds, etc.). Wright's F statistics are computed following Caballero & Toro (2000, 2002. These authors have formalized the pedigree tools necessary for the analysis of genetic differentiation in subdivided populations starting from the average pairwise coancestry coefficient (f_{ij}) between individuals of two subpopulations, i and j, of a given metapopulation including all $N_i \times N_j$ pairs. For a given subpopulation i, the average coancestry, the average self-coancestry of the N_i individuals and the average coefficient of inbreeding are, respectively, f_{ii},, s_i and F_i (so that $F_i = 2s_i{-}1$). The average distance between individuals of subpopulations i and j would be $D_{ij} = \lfloor (s_i + s_j)/2 \rfloor {-} f_{ij}$. From these parameters and the corresponding means for the entire metapopulation Caballero & Toro (2000, 2002 obtained the genetic distance between subpopulations i and j (Nei's minimum distance; Nei 1987) as $D_{ij} = \lfloor (f_{ii} + f_{jj})/2 \rfloor {-} f_{ij}$, and its average over the entire metapopulation as

$$\bar{D} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} D_{ij} N_i N_j}{N_T^2}$$

(where N_i, N_j and N_T are, respectively, the size of the corresponding populations i and j and the total population size), that are the equations (3) and (4) of Caballero & Toro (2002). Finally, the Wright's (1978) F-statistics are obtained as

$$F_{IS} = \frac{\tilde{F} - \tilde{f}}{1 - \tilde{f}}, \quad F_{ST} = \frac{\tilde{f} - \bar{f}}{1 - \bar{f}} = \frac{\bar{D}}{1 - \bar{f}},$$
$$\text{and } F_{IT} = \frac{\tilde{F} - \bar{f}}{1 - \bar{f}},$$

so that $(1{-}F_{IT}) = (1{-}F_{IS})\ (1{-}F_{ST})$, where $\tilde{f}, \tilde{F}$ are respectively the mean coancestry and the inbreeding coefficient for the entire metapopulation, and $\bar{f}$ the average coancestry for the subpopulation [see equations (3) and (6) in Caballero & Toro 2002].

At herd level, besides the effective number of herds, ENDOG computes the genetic importance of the herds in a population as the contribution of the herds with reproductive males to the population (Vassallo *et al.* 1986). Using this methodology the herds are classified as: (i) nucleus herds, if breeders use only their own males, never purchase males but sell them; (ii) multiplier herds, when breeders use purchased males and also sell males and (iii) commercial herds if they use purchased males and never sell males. Additionally, ENDOG computes the inverse of the probability that two animals taken at random in the population have their parent in the same herd for each path to know the effective number of herds supplying fathers (H_S), grandfathers (H_{SS}) and great-grandfathers (H_{SSS}) (Robertson 1953).

Finally, at population level, ENDOG computes both the generation intervals, defined as the average age of parents at the birth of their progeny kept for reproduction, and the average age of parents at the birth of their offspring (used for reproduction or not). Both parameters are computed for the four pathways (father–son, father–daughter, mother–son and mother–daughter).

Input and output files

ENDOG has been designed to avoid much need on preparation of input files. ENDOG accepts xls files (from Microsoft Excel^{TM} worksheets) or dbf files. Files with dbf format can be used in datasets larger than the limit of rows of Excel (65,536). Columns (or fields) are not supposed to be in a given order and no strict identification of the columns is needed. At the beginning of a session ENDOG will ask for a file containing the input data and, if **.xls**, for the particular worksheet in which the pedigree is. After that, the program will ask if records are renumbered and ordered sequentially (from 1 to n, the older the lower number) and, later, for the selection of the column (or field) providing the identification of the individuals, the identification of the fathers, the identification of the mothers, and the sex of the individuals. Numbering and ordering individuals is recommendable, especially if birthdates are not completely known, but, in fact, individuals can be identified in any way (using numbers, characters or both). In any case, the identifications used for individuals must be consistent with those used for parents. If records are not renumbered and sequentially ordered, ENDOG will ask for the column (or field) in which the individuals' birth date is to proceed to order data. Dates must be in dd/mm/yyyy format. Sex must be coded as 1 for males and 2 for females. Despite these shortcomings, the input file can have any other columns (or fields) in any format (character, date, numerical or other). These columns may provide any other information: different ways to identify the individuals, the identification of the herd or population corresponding to the individuals, etc. The inclusion of a column with the birth date of the animals in the input file is highly recommended because this information will be needed for some procedures. Users interested in computing parameter f_a using a particular reference population must include in the input file a column (or field) in which the animals forming the reference population are identified using a '1'.

Most results of ENDOG are written in a Microsoft ACCESS file named Gener.mdb to facilitate further use. Results of each analysis are written to the corresponding Table within Gener.mdb file. However, users may be interested in obtaining the summary results that ENDOG shows in the screen after performing some analysis. These summary results are written in their corresponding txt files with delimited pieces of information to allow their edition using any worksheet software. The names of the ACCESS tables and txt files containing the results of the computations are usually self informative on the content and are described in Table 2.

Table 2 Description of the result files obtained using ENDOG

Procedure	ACCESS table	txt results file	Description
Initial check		Error.txt	List of errors found in the pedigree
Default computations	Midef		Computes *F*, *AR*, and number of generations for each individual in the dataset
Generations Submenu	PorG PorC	Populat.txt	Mean values of *F*, *AR* and N_e for each generation traced
Founders Submenu	Ancestro RebaFund	Founders.txt Ancestor.txt	Individual and average information on ancestors explaining genetic variability and effective number of founder herds
Intervals submenu	GenInterv		Average generation intervals and reproductive ages for each path parent–son
Fstats submenu	AverDist DistNei Fis_Fsts	Coancest.txt MatFst.txt	Paired Average, Nei and Fst distance values for each defined subpopulation, and coancestry matrix
Herds structure submenu	HerStr StrHerd Roberts		Information on the genetic importance of each herd in the population, summary of these information and Robertson (1953) statistics
Coancestry submenu	Parent		Coancestry values of a key individual with all the individuals of the other sex in the dataset

Conclusions

The program ENDOG will help breeders and researchers to monitor changes in the genetic variability and structure of the populations with limited cost of preparing datasets. Although written primarily as a populations monitoring package, ENDOG does offer a number of features that may be of interest to teachers and students to develop an in-depth understanding of important statistical concepts and procedures for population genetic analysis. Despite the example file provided with the program includes a very small population, ENDOG can handle very large data files and successful computation of the parameters will be limited basically by the computer characteristics. ENDOG has been recently used to analyse 75 389 records included in the studbook of the Andalusian horse (Valera *et al.* 2005). The CPU time to obtain the complete set of computations on a PC (processor 1.8 GHz, 512 Mb RAM) was <5 min.

Acknowledgements

This paper was partially funded by two grants from INIA, no. RZ02-020 and no. RZ03-011. The authors would like to thank Luis José Royo, Isabel Álvarez and especially Iván Fernández, for their kind support and help. ENDOG has been tested on several data sets and results were checked for consistency with alternative software when possible. The authors would appreciate to be informed of any detected bug.

References

Boichard D. (2002) Pedig: a Fortran package for pedigree analysis suited for large populations. In: Y. van der Honing (ed), Proceedings of the 7th World Cong. Genet. Appl. to Livest. Prod., Montpellier, 19–23 August 2002. INRA, Castanet-Tolosan, France, CD-Rom, comm. No. 28-13.

Boichard D., Maignel L., Verrier E. (1997) The value of using probabilities of gene origin to measure genetic variability in a population. *Genet. Sel. Evol.*, **29**, 5–23.

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

Caballero A., Toro M.A. (2002) Analysis of genetic diversity for the management of conserved subdivided populations. *Conserv. Gen.*, **3**, 289–299.

Colleau J.J. (2002) An indirect approach to the extensive calculation of relationship coefficients. *Genet. Sel. Evol.*, **34**, 409–421.

Goyache F., Gutiérrez J.P., Fernández I., Gómez E., Álvarez I., Díez J., Royo L.J. (2003) Using pedigree information to monitor genetic variability of endangered populations: the Xalda sheep breed of Asturias as an example. *J. Anim. Breed. Genet.*, **120**, 95–103.

Gutiérrez J.P., Altarriba J., Díaz C., Quintanilla A.R., Cañón J., Piedrafita J. (2003) Genetic analysis of eight Spanish beef cattle breeds. *Genet. Sel. Evol.*, **35**, 43–64.

Honda T., Nomura T., Yamaguchi Y., Mukai F. (2004) Monitoring of genetic diversity in the Japanese Black cattle population by the use of pedigree information. *J. Anim. Breed. Genet.*, **121**, 242–252.

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

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

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

Nei M. (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York, pp. 512.

Quaas R.L. (1976) Computing the diagonal elements of a large numerator relationship matrix. *Biometrics*, **32**, 949–953.

Robertson A. (1953) A numerical description of breed structure. *J. Agric. Sci.*, **43**, 334–336.

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

Vassallo J.M., Díaz C., García-Medina J.R. (1986) A note on the population structure of the Avileña breed of cattle in Spain, Livest. *Prod. Sci.*, **15**, 285–288.

Woolliams J.A., Pong-Wong R., Villanueva B. (2002) Strategic optimisation of short and long term gain and inbreeding in MAS and non-MAS schemes. In: Proceedings of the 7th World Cong. Genet. Appl. to Livest. Prod., Montpellier, 19–23 August 2002. INRA, Castanet-Tolosan, France, CD-Rom, comm. No. 23_02.

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

Wright S. (1978) *Evolution and the Genetics of Populations: Vol. 4. Variability within and among Natural Populations*. University of Chicago Press, Chicago, IL, USA.