User's Guide

MolKin v3.0

A Computer Program for Genetic Analysis of Populations Using Molecular Coancestry Information

Latest update of this guide on 22-5-2009

Juan Pablo Gutiérrez, DVM, PhD[1]

Departamento de Producción Animal.
Facultad de Veterinaria.
Universidad Complutense de Madrid
Avda. Puerta de Hierro s/n
E-28040-Madrid, Spain.
Phone and Fax: +34 913943767.
E-mail: gutgar@vet.ucm.es

Félix Goyache, DVM, PhD

Área de Genética y Reproducción Animal
SERIDA-Somió
C/ Camino de los Claveles 604
E-33203 Gijón (Asturias), Spain.
Phone: +34 985195303. Fax: +34 985195310
E-mail: fgoyache@serida.org

[1]To whom correspondence is recommended

INDEX

1.- Introduction

1.1 Purpose and General Comments

MolKin (current version 3.0) is a population genetics computer program that conducts several genetic analyses on microsatellite information in a friendly user's environment. The program will help researchers or responsible for management of populations to monitor the changes in genetic variability and population structure with limited cost of preparing datasets. Moreover, although written primarily as a program for research purposes, MolKin (v3.0) does offer a number of features that may be of interest to teachers and students to develop an in-depth understanding of relevant concepts to population genetic analysis.

MolKin is tributary of some routines written for the program ENDOG (Gutiérrez and Goyache, 2005), fitted to analyse pedigree information, which is freely available at http://www.ucm.es/info/prodanim/JP_Web. MolKin has been written in VisualBasic™ 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 and from the World Wide Web at http://www.ucm.es/info/prodanim/JP_Web. MolKin 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. Despite the example file provided with the program includes a very small dataset, MolKin can handle very large data files and successful computation of the parameters will be limited basically by the computer characteristics.

1.2 Notice and Disclaimer

Please send bibliographic information, preferably a reprint, about every paper in which MolKin was used to any author of this document. Please report any errors found to author's address as indicated (preferably to Juan Pablo Gutiérrez by e-mail). We would very much appreciate users submitting their suggestions of improvements for this manual to us directly over e-mail, just sending an improved version of the MolKin User's Guide.

This program is provided 'as-is'. No authors could be held responsible in case of trouble. Although this program has been tested, the authors make no warranty as to the accuracy and functioning of the program. You may distribute this program freely in any format, so long as the following conditions are met: the program remains intact without modification, the help file is included without modification, no fee of any kind is charged.

1.3 News and Further Development

The first version of MolKin (v1.0) was successfully used in Álvarez et al. (2005). In the version 2.0, we included a bootstrapping procedure to compute, if needed, molecular coancestry and most genetic distances calculated by MolKin and the possibility of computing the Hurlbert's (1971) rarefacted number of alleles per locus. The current (v3.0) version of MolKin includes significant improvements: a) several bugs reported by the users have been solved, even though they did not affect the quality of the obtained results; b) the computation of the distance of Jürgen Tomiuk and Volker

Loeschcke (D_{TL}) in the form proposed by Tomiuk et al. (1998); c) the bootstrapping method recommended by Simianer (2002; Baumung et al., 2006) adjusting for sampling to avoid bias in estimates because of unequal populations sampling sizes; d) the methods for quantifying contribution to genetic diversity developed by Caballero and Toro (2002) and Petit et al. (1998). A bug on the computations of the internal contributions to diversity using the methodology by Petit et al. (1998) has recently been fixed. Users are kindly recommended to re-compute their data using the current version of MolKin.

Many users have suggested the inclusion in MolKin of additional parameters. Users are kindly requested to send the authors their own routines (in any programming language) with a (brief) explanation on the interest of including them in future versions of the program. These routines will be appropriately acknowledged in further modifications of this User's Guide.

The compatibility between ENDOG and Widows Vista is not completely solved. However, users can find at http://www.ucm.es/info/prodanim/html/JP_Web.htm an executable file that can be downloaded directly to be run under this environment. Also, users working under other versions of Microsoft Windows can download this executable file directly to obtain the newest version of the program to avoid the need of new installations. Also, the authors are interested in obtain a version of ENDOG running under LINUX. Users are kindly requested to suggest the most appropriate ways to deal with these tasks.

1.4 How to cite MolKin (v3.0)

If you wish to cite the use of MolKin in your publications, we suggest the following citation:

Gutiérrez, J.P., Royo, L.J., Álvarez, I., Goyache, F. (2005) MolKin v2.0: a computer program for genetic analysis of populations using molecular coancestry information. Journal of Heredity, 96: 718-721.

2.-What MolKin (v3.0) Does

Primary functions carried out by MolKin are the computation of the between individuals (and populations) molecular coancestry coefficients (f_{ij}, Caballero and Toro, 2002), the Kinship distance (D_k) at individual and population levels. Additionally, users can compute with MolKin a set of among populations, genetic distances and F-statistics (Wright, 1978) from multilocus information following Caballero and Toro (2002).

MolKin can handle genotypic input data combining various sources of information (such as microsatellittes, RFLP, allozymes or any other) with different degrees of polymorphism. User can be interested in testing the influence of these differences in the assessment of the molecular coancestry based coefficients. MolKin allows computing most variables giving the same weight to the information provided by each locus or weighting it by its Polymorphic Informative Content (PIC; Botstein et al., 1980). Results obtained by weighting by the PIC are stored in the corresponding ACCESS tables and txt files with name beginning by "W_". Additionally, users can select to compute molecular coancestry and genetic distances using bootstrapping (adjusting or not for sampling size) regardless they are weighted the PIC or not. Results are usually given at population level; however, user can compute the between-individuals

molecular coancestry and genetic distance matrices simply by clicking on a box, thus avoiding preparation of different input files. included Results obtained by bootstrapping are stored in the corresponding ACCESS tables and txt files with name finalising by "Boot".

2.1 Computation of Molecular Coancestry

Molecular coancestry between two individuals *i* and *j* at a given locus can be computed using the following scoring rules (Caballero and Toro, 2002; Eding and Meuwissen, 2001): $f_{ij,l} = \frac{1}{4}\left[I_{11} + I_{12} + I_{21} + I_{22}\right]$, where I_{xy} is 1 when allele *x* on locus *l* in individual *i* and allele *y* in the same locus in individual *j* are identical, and zero otherwise. Notice that this value can only have four values: 0, ¼, ½ and 1. The molecular coancestry between two individuals *i* and *j* (f_{ij}) can be obtained by simply averaging over *L* analyzed loci $f_{ij} = \frac{\sum_{l=1}^{L} f_{ij,l}}{L}$. The molecular coancestry of an individual *i* with itself is self-coancestry (called s_i), which is related to the coefficient of inbreeding (here homozygosity) of an individual *i* (F_i) by the formula $F_i = 2s_i - 1$. Within and between-populations molecular coancestry are simply computed averaging the corresponding values for all the within or between-population pairs of individuals.

Molecular coancestry, as formulated above, can be easily used for the analysis of genetic diversity in subdivided population (Caballero and Toro, 2002) and is related with most genetic distances used for between population studies (Eding and Meuwissen, 2001).

2.2 Average Molecular Coancestry

MolKin computes, for each analyzed individual, the average of the molecular coancestry coefficients between one individual and all the other individuals in the dataset including itself. This information is presented for each individual both within the (sub)population in which the individual is classified and for the whole population. The aim of this parameter is to ascertain the degree in which a given genotype is represented in a population thus giving additional information on the candidates to be used for reproduction in order to preserve genetic variability.

2.3 Genetic distances and F-statistics

When possible, genetic distances have been defined in terms of molecular coancestry (Eding and Meuwissen, 2001). The list of genetic distances that are computed by MolKin are:

a) at individual and population levels:

- the Kinship distance (here called D_k) between two individuals *i* and *j* is $D_k = [(s_i + s_j)/2] - f_{ij}$ (Caballero and Toro, 2002). Within and between-populations D_k is simply computed averaging the corresponding values for all the within or between-population pairs of individuals.

- the Shared Allele Distance (D_{AS}, Chakraborty and Jin, 1993) which is computed as $D_{AS} = 1 - \frac{2\bar{P}_{SAkm}}{\bar{P}_{SAk} + \bar{P}_{SAm}}$, where $\bar{P}_{SAk}$ and $\bar{P}_{SAm}$ are respectively the average proportion of shared allele between individuals belonging to population *k* and *m*, and $\bar{P}_{SAkm}$ the average proportion of shared allele between individuals belonging to populations *k* and *m*.

- the Nei's minimum distance (D_m, Nei, 1987) computed as $D_m = [(f_{kk} + f_{mm})/2] - f_{km}$, and the Nei's standard distance (D_s, Nei, 1987) computed as $D_s = -\ln [f_{km} / (f_{kk} \cdot f_{mm})^{½}]$, where f_{kk} and f_{mm} are respectively the average coancestry between individuals belonging to population *k* and *m*, and f_{km} the average coancestry between individuals belonging to populations *k* and *m*. Note that at the individual level the Nei' minimum distance coincides with D_k because the coancestry of an individual with itself is, by definition, the self-coancestry .

b) at population level

- Wright's (1969) *F*-statistics, F_{IS}, F_{ST}, and F_{IT} (defined, respectively, as heterozygote deficiency within population, heterozygote deficiency due to population subdivision and heterozygote deficiency in the total population) are obtained as $F_{IS} = \frac{\tilde{F} - \bar{f}}{1 - \bar{f}}, F_{ST} = \frac{\bar{f} - \tilde{f}}{1 - \tilde{f}}$, and $F_{IT} = \frac{\tilde{F} - \tilde{f}}{1 - \tilde{f}}$ where $\tilde{f}$, $\tilde{F}$ are respectively the mean coancestry and the inbreeding coefficient for the entire population, and $\bar{f}$ the average coancestry for the subpopulation (see Formulae (3) and (6) in Caballero and Toro, 2002). Notice that $\tilde{F}$ is not the same as genealogical inbreeding, defined as the probability that an individual has two identical alleles by descent (Malécot, 1948), but the homozygosity, referred to the identity by state.

- the Reynold's distance (D_R, Reynolds, et al., 1983) computed as $D_R = -\ln (1- F_{ST})$

- the Tomiuk and Loeschcke 's distance (D_{TL}; Tomiuk et al., 1998) computed as $D_{TL} = -\ln (I_{TL})$, where $I_{TL} = \frac{1}{r}\sqrt{\sum_{i,j} x_{ij}^{y} \sum_{i,j} y_{ij}^{x}}$, being x_{ij} and y_{ij} the frequencies of the *i*th allele at the *j*th locus within the populations *x* and *y* and *r* the number of loci. Note that, for a large number of loci, the measures I_{TL} and D_A (Takezaki and Nei, 1996), being $D_A = 1 - \frac{1}{r}\sum_{j}\sum_{i}\sqrt{x_{ij} y_{ij}}$) are closely related, with $I_{TL} \approx 1 - D_A$ (see Tomiuk et al., 1998).
Note that, using Molkin, the distance can only be computed for diploid (or haploid coding the individuals as homozygote for each loci) organisms. If you are interested in managing polyploid data please use the program POPDIST (Guldbrandtsenet al., 2000) which is freely available at http://genetics.agrsci.dk/~bernt/popgen/.

2.4 Rarefaction method

The simple average number of alleles per breed is a highly informative measure that can be useful to interpret the main results obtained using MolKin. However, this parameter can be affected by the sample size and it needs to be corrected using the Hurlbert's rarefaction method (1971) as $A[g] = \sum_i \left[1 - \prod_{k=0}^{g-1} \frac{N - N_i - k}{N - k} \right]$ where g is the specified sampled size, N the number of gene copies examined in a given locus ($N > g$) and N_i the number of occurrences if the ith allele among the N sampled gene copies.

2.5 Polymorphic Informative Content

The polymorphic informative content (*PIC*, Botstein et al., 1980) at both marker and population level is computed as $PIC = 1 - \sum_i p_i^2 - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2 p_i^2 p_j^2$, being p_i and p_j the frequency of the alleles i and j of a given locus. The parameter PIC refers to the value of a marker for detecting polymorphism within a population, depending on the number of detectable alleles and the distribution of their frequency and has been proved to be a general measure of how informative a marker is (Guo and Elston, 1999); the higher the PIC value, the more informative a marker.

2.6 Quantification of contributions to diversity

The contribution of each analysed population to the diversity of the whole dataset can be assessed using the methods proposed by Caballero and Toro (2002) and Petit et al. (1998). The former method can also be applied to an arbitrary group of populations using the same input file.

- Caballero and Toro (2002) proposed setting priorities for conservation using as criterion the maintenance of the maximum overall Nei's (1987) gene diversity (GD) in the preserved set of breeds. Note that this is equivalent to minimise the overall molecular coancestry ($\bar{f}$) because $GD = 1 - \bar{f}$. The average molecular coancestry over a entire metapopulation ($\bar{f}$) consisting of n subpopulations, subpopulation i with N_i breeding individuals, being f_{ij} be the average pairwise coancestry between individuals of subpopulations i and j, including all $N_i \times N_j$ pairs and f_{ii} the average pairwise coancestry within subpopulation i is (see formula (5) of Caballero and Toro, 2002):

$\bar{f} = \sum_{i=1}^{n} \frac{N_i}{N_T} \left[f_{ii} - \frac{\sum_{j=1}^{n} D_{ij} N_j}{N_T} \right]$, where D_{ij} is the Nei's minimum genetic distance (Nei , 1987) between subpopulations i and j (computed as $D_{ij} = [(f_{ii} + f_{jj})/2] - f_{ij}$). From this formula it is clear that average GD depends on the within-subpopulation coancestry (first term in the brackets) and the average distance among subpopulations (second term in the brackets) thus allowing to separate the contributions to the total GD due to the within breeds diversity (f_{ii}) and the between-breeds genetic distance, and $GD_T = GD_W + GD_B$, where GD_T is the total contribution to GD, GD_W is the contribution to the within-breeds diversity and GD_B the contribution to the between-breeds diversity.

Note that positive contributions to diversity from a given population using the Caballero and Toro's (2002) method mean that the remaining dataset increases the overall diversity; consequently, the assessed population would not be preferred for conservation.

- Petit et al. (1998) used the Hurlbert's (1971) rarefacted number of alleles per locus (*k*) to asses the contribution of the i^{th} population to the total allelic richness as $C_T^g(i) = \frac{\hat{k}_T^g - \hat{k}_{T\setminus i}^g}{\hat{k}_T^g - 1}$, where $\hat{k}_T^g$ is the Hurlbert's (1971) estimator of the total allelic richness in the whole analysed population, $\hat{k}_{T\setminus i}^g$ is the estimator of the total allelic richness when the i^{th} population is excluded. The partitioning of $C_T^g(i)$ in two components $C_S^g(i)$, which is the contribution to the total allelic richness due to the own allelic richness of the i^{th} population, and $C_D^g(i)$, which is the contribution due to its divergence, can be obtained as $C_S^g(i) = \frac{1}{n}\left(\frac{\hat{k}_i^g - \hat{k}_{T'\cdot i}^g}{\hat{k}_T^g - 1}\right)$, where $\hat{k}_{T'\cdot i}^g$ is the average *k* for the remaining populations after removal of population *k* and $C_D^g(i)$ simply by difference $C_D^g(i) = C_T^g(i) - C_S^g(i)$.

Note that positive contributions to diversity from a given population using the Petit et al.'s (1998) method mean that the remaining dataset has a lower number of alleles than the original one; consequently, the assessed population would be preferred for conservation.

3.- How to Use MolKin (v3.0)

Please download MolKin from the web site http://www.ucm.es/info/prodanim/JP_Web. After double-clicking on the icon of the zip file, a setup menu will guide users to install the program in the appropriate directory.

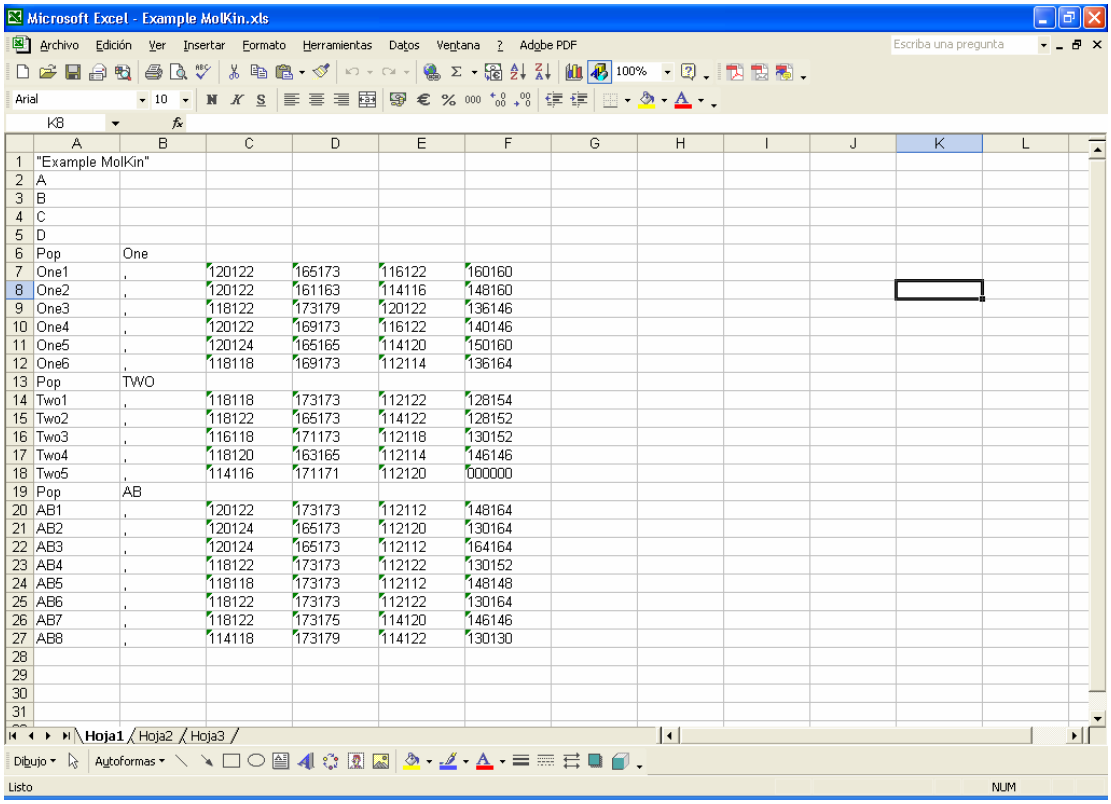
3.1 Input Files

MolKin has been thought to avoid much need on preparation of input files. MolKin accepts xls files (from Microsoft Excel worksheets) and plain text (txt) files that must contain data in GenePop (Raymond and Rousset, 1995) format with each allele coded with 3 digits. The text files can be delimited with either spaces or tabs. This format can be used to conveniently record genotypes of electrophoretic or of some microsatellite loci. The length (in nucleotides) of a microsatellite or the relative mobility of electrophoretic alleles can be directly indicated. This format makes it easier to check the input file for mistakes. Missing data are indicated as "000", as illustrated in the second populations of the above file. Note that the homozygote for the 90 allele is noted 090090 (and not 9090 as in the two digit format). In order to facilitate the use of other in put data files, MolKin permits that the two alleles of each marker are separated by a "/" (i.e. 090/090). Also, the current version of MolKin (v3.0) allows user coding missing data "999" to avoid the problems encountered by some users that can upload old .txt input files files to an Excel worksheet before starting a session with MolKin.

Hope they will be self-informative, an xls file called 'MolKin_example_input_file.xls' and a plain text file delimited by spaces ('MolKin_example_input_file.xls') are

provided with the program. Notice that MolKin can also handle plain text files delimited with tabs, and we usually work with this format.

Figure 1: 'MolKin_example_input_file.xls'



	A	B	C	D	E	F
1	"Example MolKin"					
2	A					
3	B					
4	C					
5	D					
6	Pop	One				
7	One1	,	120122	165173	116122	160160
8	One2	,	120122	161163	114116	148160
9	One3	,	118122	173179	120122	136146
10	One4	,	120122	169173	116122	140146
11	One5	,	120124	165165	114120	150160
12	One6	,	118118	169173	112114	136164
13	Pop	TWO				
14	Two1	,	118118	173173	112122	128154
15	Two2	,	118122	165173	114122	128152
16	Two3	,	116118	171173	112118	130152
17	Two4	,	118120	163165	112114	146146
18	Two5	,	114116	171171	112120	000000
19	Pop	AB				
20	AB1	,	120122	173173	112112	148164
21	AB2	,	120124	165173	112120	130164
22	AB3	,	120124	165173	112112	164164
23	AB4	,	118122	173173	112122	130152
24	AB5	,	118118	173173	112112	148148
25	AB6	,	118122	173173	112122	130164
26	AB7	,	118122	173175	114120	146146
27	AB8	,	114118	173179	114122	130130

3.2 Output Files

Most results of MolKin are written in a Microsoft ACCESS file named Microsat.mdb to facilitate further use. Results of each analysis are written to the corresponding Table within Microsat.mdb file. However, user can be interested in obtaining the summary results that MolKin shows in the screen after performing some analysis. These summary results are written in their corresponding ACCESS Tables and txt files. The latter are 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. In any case, these are summarized in Table 1.

Regardless most parameters are computed at population level, MolKin has been programmed to give some results at individual level (usually in plain text results file) using the same input file. User, however, can always rearrange the input file fitting as many populations as individuals in the file to obtain genetic parameters at individual level in the corresponding ACCESS Table.

Table 1: List of the result files obtained using MolKin

Submenu	ACCESS table	txt result files	Description
Default computations	MicroSat DiverInd (W_)[1]		Computes the number of alleles, observed heterozygosity and the Polymorphic Informative Content (PIC) for each analyzed marker; additionally it gives the number of copies of each

			allele in the dataset and the corresponding intralocus allelic frequencies. On the other hand, the DiverInd Table gives the proportion of diversity between individuals for subpopulation *i*.
Distances Consult	DistancesSCREEN KinDist (W_)[1] DistMinNei (W_) DistStdNei (W_) DReynold (W_) Fis_Fsts (W_) Kinsub (W_) SADSub TomLoe	MatFst.txt (W_) Mat(non0).txt (W_)[2] MAvD(N0).txt (W_)[2] (SAD_Mat).txt[2]	First, the DistancesSCREEN Table gives the average values that are shown in the main screen of the Distances Menu. The other Tables give the between population distance matrix expressed in the corresponding name. KinDist corresponds to the kinship distance (D_k) and Kinsub corresponds to the molecular coancestry matrix. Fis values and whitin population molecular coancestry are on diagonals of the corresponding Tables listed on the left. The txt file MatFst.txt gives the values of Fst, Fis and Fit for the whole population. The SADSub and the TomLoe tables include the between population shared allele distance matrix and the distance by Tomiuk and Loeschcke, respectively. The (SAD_Mat).txt includes the between individuals shared allele distance matrix in txt format. Additionally user can obtain the between individuals molecular coancestry matrix and the between individuals kinship distance matrix in txt format.
Distances Bootstrapping	Boots_SCREEN DiverInBoot (W_)[1] KinDistBoot (W_)[1] DistMinNeiBoot (W_) DistStdNeiBoot (W_) DReynoldBoot (W_) Fis_FstsBoot (W_) KinsubBoot (W_) SADSubBoot TomLoeBoot	MatFstBoots.txt W_MatFstBoots.txt	This submenu gives similar results than the Distances-Consult submenu but using bootstrapping. The Boots_SCREEN Table gives the average values (with their standard errors) obtained by bootstrapping of the same parameters showed in the DistancesSCREEN Table and the other Tables and txt files are equivalent to those obtained with similar calling in the Distances-Consult submenu.
Equalized size	Equal_SCREEN DiverInSamp (W_)[1] KinDistSamp (W_)[1] DistMinNeiSamp (W_) DistStdNeiSamp (W_) DReynoldSamp (W_) Fis_FstsSamp (W_) KinsubSamp (W_) SADSubSamp TomLoeSamp	MatFstSamp.txt W_MatFstSamp.txt	This submenu gives similar results than the Distances-Consult submenu but using bootstrapping equaling for sampling size. The Equal_SCREEN Table gives the average values (with their standard errors) obtained by bootstrapping equaling for sampling size of the same parameters showed in the DistancesSCREEN Table and the other Tables and txt files are equivalent to those obtained with similar calling in the Distances-Consult submenu.
Gain/Loss	GainLoss		This submenu gives the results of quantification of contribution to diversity following caballero and Toro (2002) and Petit et al. (1998)
Individual Kinship	MoleMeanKinship		This Table gives the identification of each individual and its population, and the average of the molecular coancestry coefficients of each individual with all the others in the analysed dataset. This parameter is given for the metapopulation and for the population into which each individual is assigned (unweighted and weighted by the PIC)
Individual Distances		I_A_MNei.txt I_A_SNei.txt I_W_MNei.txt I_W_SNei.txt I_TomLoe.txt	This submenu allows obtaining the between-individuals Nei's minimum, Nei's standard and Tomiuk and Loeschcke's distance matrices in plain .txt files. The Nei's minimum and Nei's standard distance matrices can also be obtained weighted by the PIC.
Rarefaction method	Rarefaction		This Table gives the observed heterozygosity, the

average PIC value, the average number of allele per locus and the average number of allele per locus using the rarefaction method

[1]A "(W_)" means that results obtained weighting by the PIC value are provided. In these cases the "W_" precedes to the name of the corresponding output file or table

[2]These matrices are provided in different formats to facilitate user to capture them. Those in which individual labels begin with a "#" are prepared to be captured with limited changes with the MEGA4 software (Tamura et al., 2007) and compute a phylogenetic tree.

As stated above, denomination and structure of the ACCESS tables and txt files including the results of the computations carried out with MolKin are self-informative. As an example please see the Figure 2 containing the Fis_Fst Table: The figures on diagonal are the Fis values for the corresponding population whilst figures below diagonal are the between-population Fst values. This structure is the same for the KinDist and the Kinsub tables but including on diagonal the within-population D_k and f_{ii} values, respectively. No other within-population genetic distances are stored in the corresponding ACCESS Table.

Figure 2: Fis_Fsts Table obtained from theMicrosat.mdb results file

Fis_Fsts : Tabla (Vista preliminar)

Fis_Fsts	ONE	TWO	THREE
ONE	0,1020408		
TWO	0,1266866	0,002214499	
THREE	0,1473613	0,1242333	0,04191617

The results Table corresponding Figure 2 when bootstrapping is used (regardless it adjusts for sampling size or not) is shown in Figure 3. Notice that the standard errors are below the corresponding value. The number of samples on which bootstrapping has been carried out is always at the bottom of the Table.

Figure 3: Fis_FstsBoots Table obtained from theMicrosat.mdb results file

Fis_FstsBoots : Tabla

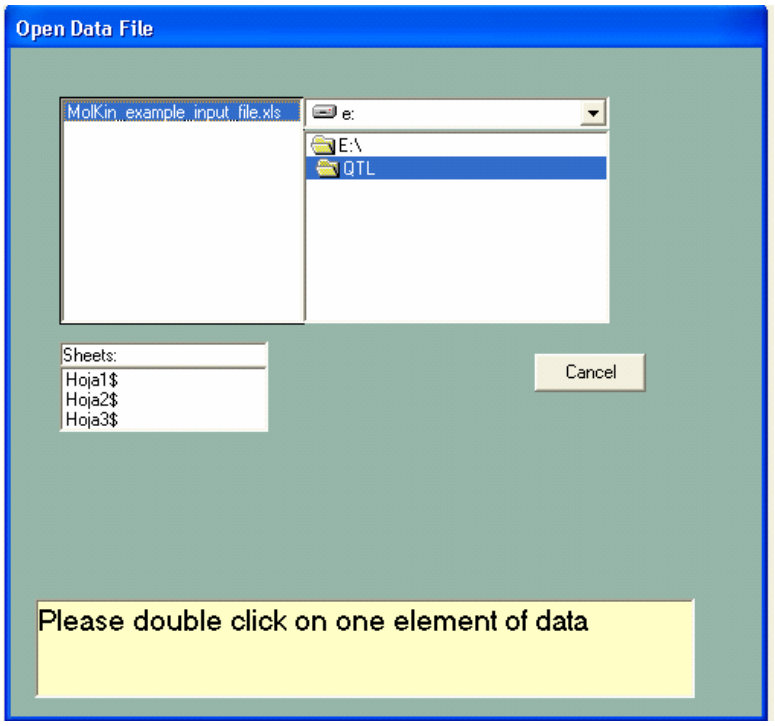
Fis_Fsts	ONE	TWO	THREE
ONE	0,05763323		
ONE_STD	0,07317125		
TWO	0,1442692	-0,03764075	
TWO_STD	0,02831478	0,09214904	
THREE	0,1704085	0,1487754	-0,004225795
THREE_STD	0,02142042	0,02240837	0,08482743
Nº Samples =	100		

The other Tables and txt files containing the results of the computation carried out with MolKin are expected to be self-informative. In any case a brief inspection of the information summarized in table 1 can be useful for the first approach to MolKin.

3.3 A Session with MolKin (v3.0)

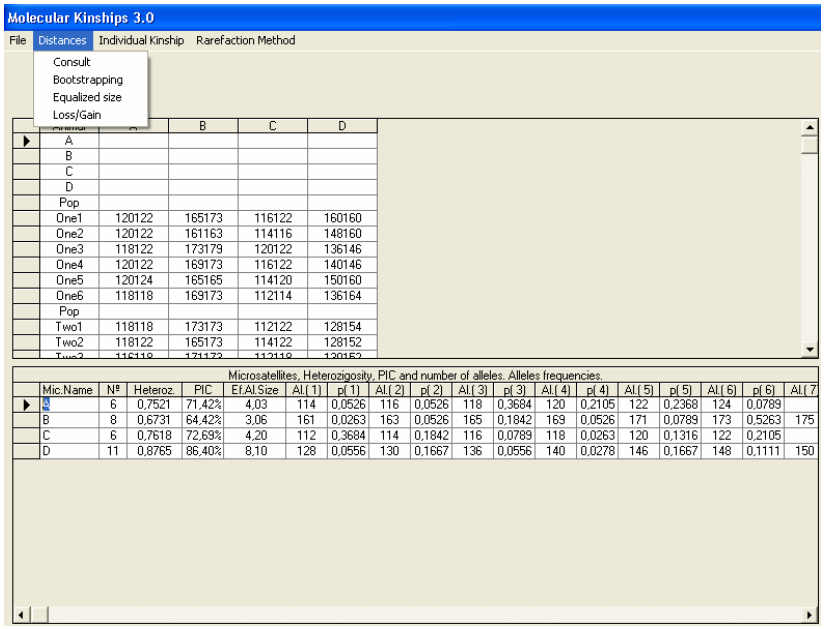
MolKin starts by simply double-clicking on the icon of the program. After that user can see the Initial screen of MolKin (Figure 5); this screen allows user to find the input data file in the corresponding directory. After double-clicking on the selected xls file MolKin will permit to select the worksheet on which the session will be carried out.

Figure 4: Initial screen of MolKin



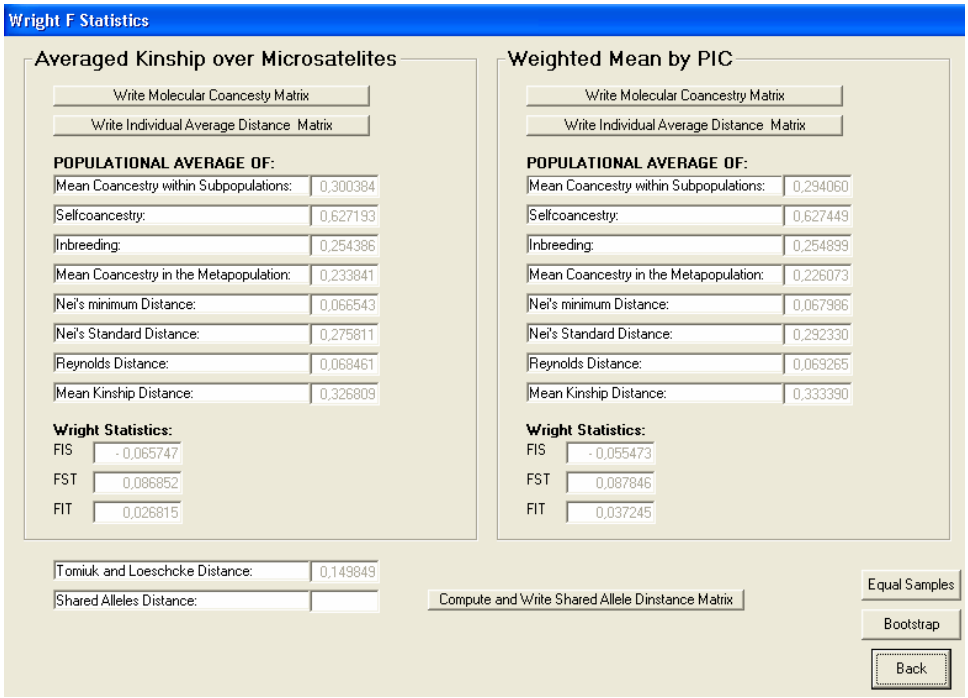
After loading the input file some default computations are done. The main screen of MolKin will give two windows (Figure 5). In the upper window the user can observe the default computation done whilst the other window shows the input data. Notice that if the input data are in txt format the user will only see the default computations. In the Menu bar, user will find three different menus: Distances, Individual Kinship and Rarefaction.

Figure 5: Main screen of MolKin



Clicking on the Distances Menu user will find four different submenus: a) the Consult submenu; b) the Bootstrap sumneu; c) the Equal size submenu; and d) the gain/Loss submenu. Clicking on the Consult submenu, user will find the main statistics computed for the whole dataset and, clicking in the corresponding box of the screen (see Figure 6), write to disk the between individuals molecular coancestry matrix, the between individuals average (kinship) distance matrix and the between individuals shared allele distance matrix. User can obtain most parameters using bootstrapping (regardless it adjust for sampling size or not) without need of obtaining previously the direct results using the Bootstrapping or the Equal size submenus. However, one can always obtain the bootstrapping estimates simply clicking on the Bootstrap or the Equal Samples buttons included in the screen of the Distances-Consult submenu.

Figure 6: Screen of Distances menu



After clinking on the Distances-Bootstrapping submenu or on the Distances-Equal size submenu (or on the corresponding buttons of the screen of the Distances-Consult submenu) a dialog box will appear (Figures 7 and 8). In this box user can select to compute one or more of the following parameters: a) the coancestry-related genetic parameters without weighting them by the PIC (*Averaged Kinship over Microsatellites*); b) the coancestry-related genetic parameters weighting them by the PIC (*Averaged Kinship over Microsatellites*); c) the Shared Allele Distance; and d) the Tomiuk and Loeschcke's distance. Moreover, user can fit the number of bootstrapping samples (and the number of samples per population; see Figure 8). These dialog boxes allow users to decide how much time they would like to spend during the computations. Notice that bootstrapping can be both time and computer consuming if the number of samples fitted and the data set are large.

After clicking on the Gain/Loss submenu the contributions of the analysed populations to the total diversity computed using the methods by Caballero and Toro (2002) and Petit et al. (1998) will appear. The Table GainLoss includes the following fields: GD, Internal Diversity, Mean Distance, Loss/Gain (which include, respectively, the Genetic Diversity of the dataset after removing the population, the within-population contribution to GD, the he between-populations contribution to GD – which is the Nei' minimum distance- and the total contribution to GD), Pe_Int_Diversity, Pe_Divergence and Petit(*g*)% (which include, respectively, within-population, between-populations and total contributions for diversity using the method by Petit et al. (1998), being *g* the number of copies used for rarefaction. All the values are given in percentage.

Figure 7: Dialog box of the Boostrapping procedure

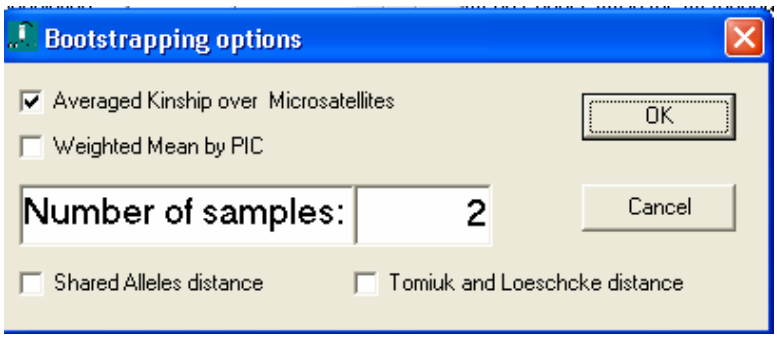


Figure 8: Dialog box of the Equal size procedure

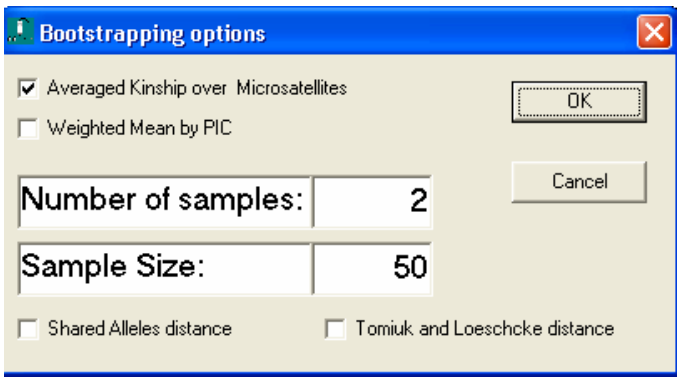
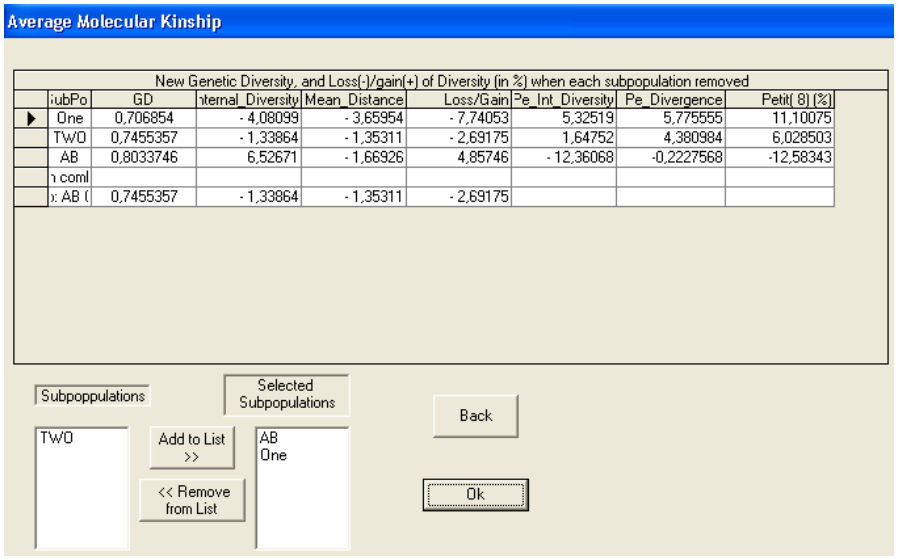


Figure 9: Screen of the Gain/Loss procedure



The second Menu of MolKin is the Individual Kinship Menu. It has three submenus: a) the Between Individuals submenu; b) the Mean Molecular Kinship submenu; and c) the Individual Distances submenu. The Between Individuals submenu has been thought to help breeders in the management of a given population: user selects any couple of individuals to be mated and obtain the corresponding molecular coancestry coefficients. To facilitate the interpretation of the results the average values for the whole analysed population are presented in the screen (Figure 10). The Mean Molecular Kinship submenu (Figure 11) computes the average molecular coancestry between each individual and all the others in the whole population or in the subpopulation in which the individual is classified (unweighted or weighted by the PIC). The Individual Distances submenu (see Figure 12) allows obtaining some between-individuals distance matrices (unweighted or weighted by the PIC).

Figure 10: Screen of MolKin for the Between Individuals submenu

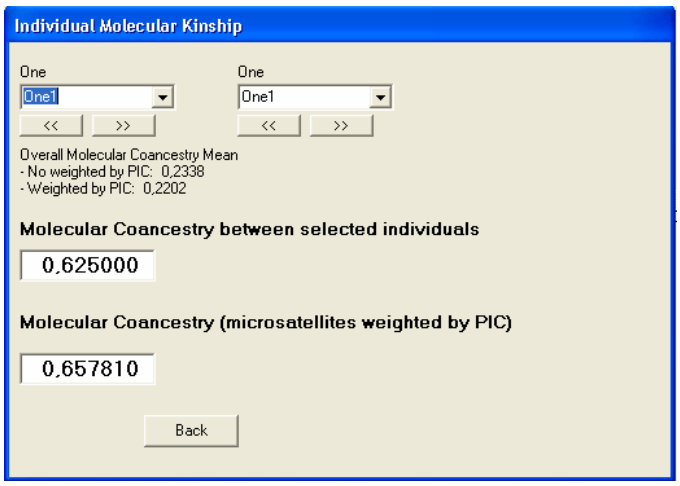


Figure 11: Screen of MolKin for the Mean Molecular Kinship submenu

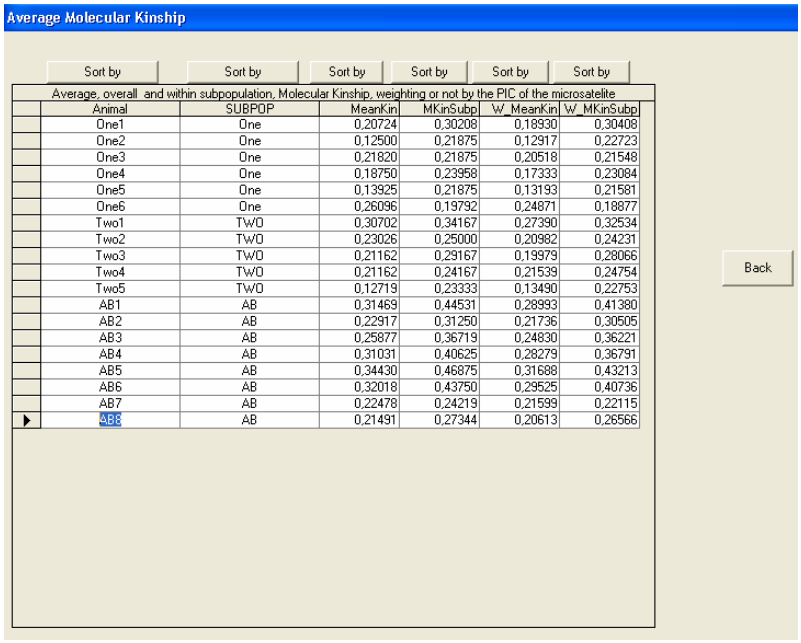
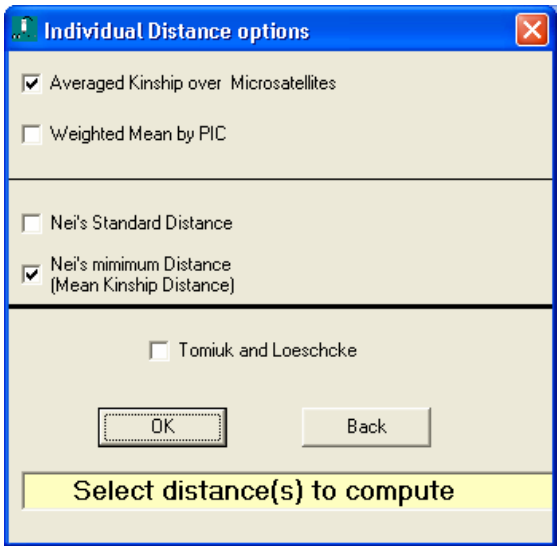
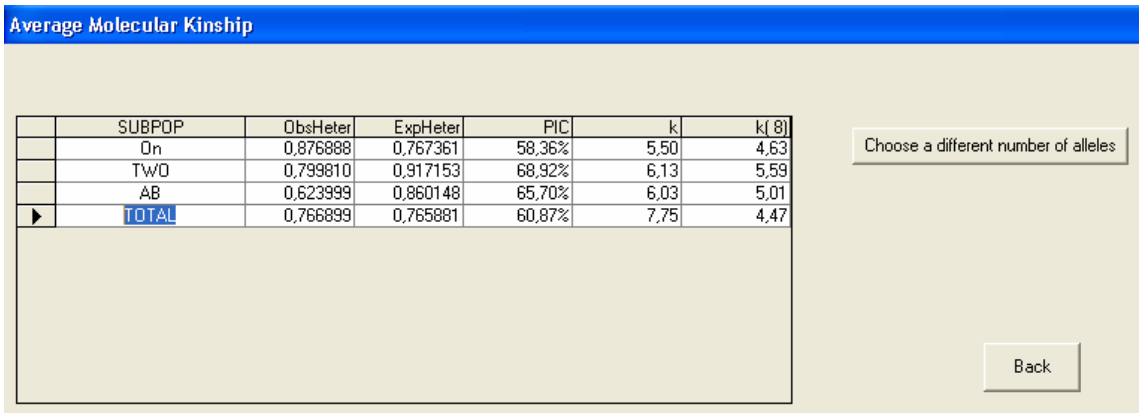

Average Molecular Kinship
Sort by
Sort by
Sort by
Sort by
Sort by
Sort by
Average, overall and within subpopulation, Molecular Kinship, weighting or not by the PIC of the microsatelite
Animal SUBPOP MeanKin MKinSubp W_MeanKin W_MKinSubp
One1 One 0,20724 0,30208 0,18930 0,30408
One2 One 0,12500 0,21875 0,12917 0,22723
One3 One 0,21820 0,21875 0,20518 0,21548
One4 One 0,18750 0,23958 0,17333 0,23084
One5 One 0,13925 0,21875 0,13193 0,21581
One6 One 0,26096 0,19792 0,24871 0,18877
Two1 TWO 0,30702 0,34167 0,27390 0,32534
Two2 TWO 0,23026 0,25000 0,20982 0,24231
Two3 TWO 0,21162 0,29167 0,19979 0,28066
Two4 TWO 0,21162 0,24167 0,21539 0,24754
Two5 TWO 0,12719 0,23333 0,13490 0,22753
AB1 AB 0,31469 0,44531 0,28993 0,41380
AB2 AB 0,22917 0,31250 0,21736 0,30505
AB3 AB 0,25877 0,36719 0,24830 0,36221
AB4 AB 0,31031 0,40625 0,28279 0,36791
AB5 AB 0,34430 0,46875 0,31688 0,43213
AB6 AB 0,32018 0,43750 0,29525 0,40736
AB7 AB 0,22478 0,24219 0,21599 0,22115
AB8 AB 0,21491 0,27344 0,20613 0,26566
Back

Figure 12: Screen of MolKin for the Individual Distance submenu



The third menu of MolKin is the Rarefaction Method Menu (Figure 13). It gives an ACCESS Table showing, for each population, the observed heterozygosity, the expected heterozygosity, the average PIC, the average number of observed alleles per locus (A) and the average number of alleles per locus corrected using the rarefaction method (A(g)) where n is the normalized allele size of the population, depending on the lower population size in the dataset. The population size is computed considering only those individuals with known values for all the genotyped markers. Notice that g = 8 means that the lower number of individuals with full genotypic information in a population is 4. User can fit a particular value of n clicking on the 'Choose a different number of alleles' button. Of course, the new value for n can not exceed that fitted by default.

Figure 13: Screen of MolKin for the Rarefaction Method Menu



4.- References

Álvarez I., Royo L.J, Fernández I., Gutiérrez, J.P.,Gomez, E., Arranz, J.J., Goyache, F. (2005) Testing the usefulness of the molecular coancestry information to assess genetic relationships on livestock using a set of Spanish sheep breeds. *Journal of Animal Science*, in press.

Baumung, R., Cubric-Curik, V., Schwend, K., Achmann,R., Sölkner, J., 2006. Genetic characterisation and breed assignment in Austrian sheep breeds using microsatellite marker information. J. Anim. Breed. Genet. 123: 265-271.

Botstein D., White R.L., Skolnick M., Davis R.W (1980) Construction of a genetic linkage map in man using restriction fragment polymorphisms. *Am J Hum Genet* 32: 314-331.

Caballero A, Toro MA, 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 MA, 2002. Analysis of genetic diversity for the management of conserved subdivided populations. Conserv Gen 3: 289-299.

Chakraborty R, Jin L (1993) Determination of relatedness between individuals using DNA fingerprinting. Hum Biol. **65**: 875-95.

Eding H., Meuwissen, THE. (2001) Marker-based estimates of between and within population kinships for the conservation of genetic diversity. Journal of Animal Breeding and Genetics 118**,** 141-59.

Guldbrandtsen, B., Tomiuk, J., Loeschcke, V. (2000) POPDIST, "Version 1.1.1: A Program to Calculate Population Genetic Distance and Identity Measures", J. Hered. 91(2), 178-179.

Guo, X., and Elston, R.C. (1999) Linkage informative content of polymorphic genetic markers. Hum. Hered. 49:112-118.

Gutiérrez, J.P., Goyache F. (2005) A note on ENDOG: a computer program for analysing pedigree information. J. Anim. Breed. Genet., 122: in press.

Hurlbert, S.H., 1971. The non concept of species diversity: a critique and alternative parameters. Ecology 52: 577-586.

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

Petit, R.J., El Mousadik, A., Pons, O., 1998. Identifying populations for conservation on the basis of genetic markers. Conserv. Biol. 12, 844–855.

Raymond, M., Rousset, F. 1995. GENEPOP (Version 1.2): Populations genetic software for exact test and ecumenicism. J. Hered. 86:248-249.

Reynolds, J., Weir, B.S., Cockerham, C. 1983. Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics, 105, 767-779,

Simianer H., 2002. Molekulargenetische Differenzierung verschiedener Rotviehpopulationen. Schriftenreihe des Bundesministeriums für Verbraucherschutz, Ernährung und Landwirtschaft. Heft 493. Landwirtschaftsverlag GmbH. Münster-Hiltrup, Germany.

Takezaki, N., Nei, M., 1996. Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. Genetics 144: 389–399.

Tamura K, Dudley J, Nei M, Kumar S, 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24, 1596-1599.

Tomiuk, J., Guldbrandtsen, B., Loeschcke, V., 1998. Population differentiation through mutation and drift – a comparison of genetic identity measures. Genetica 102/103: 545–558.

Wright S, 1978. Evolution and the genetics of populations: Vol. 4. Variability within and among natural populations. University of Chicago Press: Chicago. USA

5.- Acknowledgements

This document has been developed in a collaborative effort in which has been involved the Breeders Association of Xalda Sheep of Asturias (http:/www.xalda.com/) and the breeders association of Poni Asturcón within conservation efforts partially funded by grants from INIA, no. RZ02-020 and RZ03-011. Some modifications included in the version v3.0 of the program have been done in the framework of the project MEC-FEDER CGL2005-03761/BOS. The authors would like to thank Luis J. Royo, Isabel Álvarez and especially Iván Fernández, for their kind support and help.

All of the comments and suggestions on MolKin have been greatly appreciated; Ino Curik showed a gratifying interest on the program and suggested the inclusion of several improvements in the program; Dr Paul Johnson of the DEEB of the University of Glasgow has kindly informed on a bug on the reading of the .txt input files. Authors are indebted to Henner Simianer and Roswitha Baumung for providing the FORTRAN routines for bootstrapping equalling for sampling size that have been included in the current version of the program MolKin. Félix Goyache is grateful to Jürgen Tomiuk for his patience explanations on the nice properties of the D_{TL} distance. Mikhail Ozerov and Akarapong Swatdipong, from the University of Turku, have has kindly informed on a bug on the computations of the internal contributions to diversity using the methodology by Petit et al. (1998).

6.- Published papers using MolKin

Álvarez I., Gutiérrez, J.P., Royo L.J, Fernández I., Gomez, E., Arranz, J.J., Goyache, F. (2005) Testing the usefulness of the molecular coancestry information to assess genetic relationships on livestock using a set of Spanish sheep breeds. *Journal of Animal Science*, 83: 737-744.

Alfonso. L., Parada, A., Legarra, A., Ugarte, E., Arana, A. (2006) The effects of selective breeding against scrapie susceptibility on the genetic variability of the Latxa Black-Faced sheep breed. *Genetics Seletion Evolution*, 38: 495-511

Gutiérrez-Gil B., Uzun M., Arranz J.J., San Primitivo F., Yildiz S., Cenesiz M., Bayón Y. (2006) Genetic diversity in Turkish sheep. *Acta Agriculturae Scand Section A*, 56: 1-7.

Johnson, P.C.D., Webster, L.M.I., Adam, A., Buckland, R., Dawson, D.A., Keller, L.F. (2006) Abundant variation in microsatellites of the parasitic nematode *Trichostrongylus*

tenuis and linkage to a tandem repeat. *Molecular and Biochemical Parasitology*, 148: 210-218.

Royo, L.J., Pajares, G., Álvarez, I., Fernández, I., Goyache F. (2007) Genetic variability and differentiation in the Spanish roe deer (Capreolus capreolus): a phylogeographic reassessment in the European framework. *Molecular Phylogenetics and Evolution*, 42: 47-61.

Druml, T., Curik, I., Baumung, R., Aberle, K., Distl, O., Sölkner, J. (2007) Individual-based assessment of population structure and admixture in Austrian, Croatian and German draught horses. *Heredity*, 98: 114 – 122, doi:10.1038/sj.hdy.6800910

Kakoi, H., Tozaki, T., Gawahara, H. (2007) Molecular Analysis Using Mitochondrial DNA and Microsatellites to Infer the Formation Process of Japanese Native Horse Populations. *Biochemical Genetics*, 45: 375-395. doi:10.1007/s10528-007-9083-0

Ciampolini, R., Cecchi, F., Mazzanti, E., Ciani, E., Tancredi, M., De Sanctis, B. (2007) The genetic variability analysis of the Amiata donkey breed by molecular data. *Italian Journal of Animal Science*, 6 (SUPPL. 1): 78-80.

Royo, L.J., Álvarez, I., Gutiérrez, J.P., Fernández, I., Goyache, F. (2007) Genetic variability in the endangered Asturcón pony assessed using genealogical and molecular information. *Livestock Science*, 107: 162-169. doi:10.1016/j.livsci.2006.09.010

Álvarez, I., Royo, L.J., Gutiérrez, J.P., Fernández, I., Arranz, J.J., Goyache, F. (2007) Genetic diversity loss due to selection for scrapie resistance in the rare Spanish Xalda sheep breed. *Livestock Science*, 111: 204–212. doi:10.1016/j.livsci.2007.01.147

Glowatzki-Mullis, M.-L., Muntwyler, J., Bäumle, E., Gaillard, C. (2008) Genetic diversity measures of Swiss goat breeds as decision-making support for conservation policy. *Small Ruminant Research*, 74: 202-211

Álvarez, I., Royo, L. J., Gutiérrez, J. P., Fernández, I., Arranz, J. J., Goyache F. (2008) Relationship between genealogical and microsatellite information characterising losses of genetic variability: empirical evidence from the rare Xalda sheep breed. *Livestock Science*, 115: 80-88. doi:10.1016/j.livsci.2007.06.009

Legaz, E., Álvarez, I., Royo, L.J., Fernández, I., Gutiérrez, J.P., Goyache, F. (2008) Genetic relationships between Spanish Assaf (Assaf.E) and Spanish native dairy sheep breeds. *Small Ruminant Research,* 80: 39-44. doi:10.1016/j.smallrumres.2008.09.001

Dalvit, C., Sacca, E., Cassandro, M., Gervaso, M., Pastore, E., Piasentier, E. (2008) Genetic diversity and variability in Alpine sheep breeds. *Small Ruminant Research*, 80: 45-51.

Thirstrup, J.P., Pertoldi, C., Loeschcke, V. (2008) Genetic analysis, breed assignment and conservation priorities of three native Danish horse breeds. *Animal Genetics*, 39: 496-505.

d'Angelo, F., Albenzio, M., Sevi, A., Ciampolini, R., Cecchi, F., Ciani, E., Muscio, A. (2009) Genetic variability of the Gentile di Puglia sheep breed based on microsatellite polymorphism. *Journal of Animal Science*, 87: 1205-1209.

Glowatzki-Mullis, M.-L., Muntwyler, J., Bäumle, E., Gaillard, C. (2009) Genetic diversity measures of Swiss sheep breeds in the focus of conservation research. *Journal of Animal Breeding and Genetics*, 126: 164-175.

Traoré, A., Álvarez, I., Tamboura, H.H., Fernández, I., Kaboré, A., Royo, L.J., Gutiérrez, J.P., Ouédraogo-Sanou, G., Sawadogo, L., Goyache, F. (2008) Genetic characterisation of Burkina Faso goats using microsatellite polymorphism. Livestock Science, 123, 322-328.

Álvarez, I., Traoré, A., Tamboura, H.H., Kaboré, A., Royo, L.J., Fernández, I., Ouédraogo-Sanou, G., Sawadogo, L., Goyache, F. (2008) Microsatellite analysis characterizes Burkina Faso as a genetic contact zone between Sahelian and Djallonké sheep. *Animal Biotechnology*, 20: 47-57.