Journal of Heredity 2005:96(6):718–721
doi:10.1093/jhered/esi118
Advance Access publication October 26, 2005

Computer Notes

MolKin v2.0: A Computer Program for Genetic Analysis of Populations Using Molecular Coancestry Information

J. P. GUTIÉRREZ, L. J. ROYO, I. ÁLVAREZ, AND F. GOYACHE

From the Departamento de Producción Animal, Facultad de Veterinaria, Avda. Puerta de Hierro s/n, E-28040-Madrid, Spain (Gutiérrez); and SERIDA-CENSYRA-Somió, C/ Camino de los Claveles 604, E-33203 Gijón (Asturias), Spain (Royo, Álvarez, and Goyache).

Address correspondence to J.P. Gutiérrez at the address above, or e-mail: gutgar@vet.ucm.es.

Recently different studies have formalized the way in which it is possible to obtain coancestry coefficients from molecular information (Caballero and Toro 2002; Eding and Meuwissen 2001) by applying Malécot's (1948) definition of kinship to marker genes, though referring it to identity-by-state instead of identity-by-descent (Caballero and Toro 2002). The molecular coancestry between two individuals, *i* and *j*, is the probability that two randomly sampled alleles from the same locus in two individuals are identical by state. Because of its straightforward relationship with genealogical coancestry, this parameter has been shown to have interesting properties that may be used for conservation purposes (Eding et al. 2002; Toro et al., 2002; 2003). Moreover, molecular coancestry can be used to assess genetic diversity within and between populations (Eding and Meuwissen 2001). Using simulated data, Eding and Meuwissen (2001) showed that molecular coancestry has some interesting properties, namely that average kinship between populations becomes constant very quickly after population fission, causing between-population diversity to remain constant. This property allows researchers using molecular coancestry information to study the genetic relationships between populations (Álvarez et al. 2005; Caballero and Toro 2002; Fabuel et al. 2004).

Despite the utility of molecular coancestry for conservation worth and evolutionary studies, no computer routines are available to facilitate the use of molecular coancestry information. MolKin (version 2.0) is a population genetics computer program that conducts several genetic analyses on multilocus information in a user-friendly environment. The program will help researchers or those responsible for population management to assess genetic variability and population structure at reduced costs with respect to dataset preparation. A previous version of MolKin (version 1.0) was available on request for research purposes (Álvarez et al. 2005). Following Bennewitz and Meuwissen (2005), who have recently suggested that bootstrapping could significantly improve kinship estimates, the main change included in the present version of MolKin (version 2.0) is the inclusion of a bootstrapping procedure to compute, when needed, molecular coancestry coefficients and most genetic distances calculated by MolKin with the corresponding standard errors. Following Felsenstein (1985), the bootstrapping procedure implemented in MolKin (version 2.0) involves creating new datasets by randomly sampling individuals with replacement, so that the resulting datasets have the same size as the original, but some genotypes have been left out and others are duplicated. The random variation of the results from analyzing these bootstrapped datasets can be shown statistically to be typical of the variation that you would get from collecting new datasets.

Although written primarily as a program for research purposes, the new version of MolKin (version 2.0) improves the user's environment and offers a number of features that may be of interest to teachers and students for developing an in-depth understanding of concepts related to population genetic analysis.

Program Functions

The primary functions carried out by MolKin are the computation of between-individual (and population) molecular coancestry coefficients (f_{ij}) (Caballero and Toro 2002) and the kinship distance (D_k) at individual and population levels. Using MolKin, the user can also compute a set of among-population genetic distances and *F*-statistics (Wright 1978) from multilocus information. The molecular coancestry between two individuals, *i* and *j*, is the probability that two randomly sampled alleles from the same locus in two individuals are identical by state (Caballero and Toro 2002). This can be computed at a given locus using the following scoring rules (Caballero and Toro 2002; Eding and Meuwissen 2001):

$$f_{ij,l} = \tfrac{1}{4}[I_{11} + I_{12} + I_{21} + I_{22}],$$

where I_{xy} is 1 when allele X on locus *l* in individual *i* and allele *y* on 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, $f_{ij} = \left(\sum_{l=1}^{L} f_{ij,l}\right)/L$. The molecular coancestry of an individual *i* with itself is self-coancestry (called s_i), which is related to the coefficient of inbreeding of an individual *i* (F_i) by the formula $F_i = 2s_i - 1$. In turn, 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). MolKin computes within- and between-breed molecular coancestry and D_k by simply averaging the corresponding values for all the within- or between-population pairs of individuals.

Molecular coancestry is related to the majority of genetic distances used for between-population studies and F-statistics (Caballero and Toro 2002; Eding and Meuwissen 2001). Since these parameters are widely used in genetic studies (Álvarez et al. 2005; Takezaki and Nei 1996; Tomiuk et al. 1998), their computation has been implemented in MolKin (version 2.0). However, unlike other available programs such as GENEPOP (Raymond and Rousset 1995) or Fstat (Goudet 1995), MolKin computes these parameters in the following way:

(1) Wright's (1978) F-statistics—F_IT, F_ST, and F_IS—are obtained as

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

where $\tilde{f}$, $\tilde{F}$ are, respectively, the mean coancestry and the inbreeding coefficient for the whole population, and $\bar{f}$ is the average coancestry for the subpopulation [see Equations (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 homozygosity, which refers to identity by state.

(2) Nei's minimum distance (D_m) and Nei's standard distance (D_s) (Nei 1987), computed as $D_\mathrm{m} = [(f_{kk} + f_{mm})/2] - f_{km}$ and $D_\mathrm{s} = -\ln[f_{km}/(f_{kk}f_{mm})^{½}]$, respectively, where f_{kk} and f_{mm} are the average coancestry between individuals belonging to population k and m, and f_{km} is the average coancestry between individuals belonging to populations k and m.

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

MolKin further computes, at individual and population levels, 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 the average proportion of shared alleles between individuals belonging to populations k and m, and $\bar{P}_{SAkm}$ is the average proportion of shared alleles between individuals belonging to populations k and m.

Finally, MolKin computes the rarefaction method (Hurlbert 1971) and the polymorphic informative content (PIC) (Botstein et al. 1980). The simple average number of alleles (or allelic richness) per population is a highly informative measure that can be useful in interpreting the main results obtained using MolKin. However, this parameter can be affected by the sample size and needs to be corrected using Hurlbert's rarefaction method (1971) as

$$A[g] = \sum_i \left[1 - \frac{\begin{bmatrix} N - N_i \\ g \end{bmatrix}}{\begin{bmatrix} N \\ g \end{bmatrix}} \right],$$

where $A[g]$ is the rarefacted allelic richness, g is the specified sampled size, N is the number of gene copies examined in a given locus ($N > g$), and N_i is the number of occurrences of the ith allele among the N sampled gene copies. On the other hand, the PIC (Botstein et al. 1980) at both the marker and population level is computed as $PIC = 1 - \sum_i p_i^2 - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2p_i^2 p_j^2$; p_i and p_j being 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 proven to be a general measure of how informative a marker is (Guo and Elston 1999); the higher the PIC value, the more informative the marker.

Input and Output Files

MolKin has been designed to avoid the need for a large amount of preparation of input files. MolKin accepts plain text files or .xls files (from Microsoft Excel worksheets), which must contain data in GENEPOP (Raymond and Rousset 1995) format, with each allele coded with three digits. This format can be used to conveniently record genotypes of electrophoretic or 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." Note that the homozygote for the 90 allele is denoted as 090090 (and not 9090, as in the two-digit format). However, MolKin allows the two alleles of each marker to be separated by a forward slash (i.e., 090/090).

Most of the results of each analysis carried out using MolKin are written to the corresponding table in a Microsoft Access file called Microsat.mdb to facilitate their further use. In addition, MolKin has been programmed to give some results (usually at an individual level) in plain text files with tab or space delimited items of information, thus allowing their subsequent editing using any worksheet software. Moreover, some of the plain text files, including the between-individual molecular coancestry, D_k, and D_AS matrices, are provided with format prepared to be captured with limited changes using the MEGA2 software (Kumar et al. 2001) so that a phylogenetic tree can be computed. Since MolKin can handle genotypic input data that is a combination of various sources of information (such as microsatellites, restriction fragment length polymorphism [RFLP], allozymes, or any others) with different degrees of polymorphism, the user may be interested in testing the influence of these differences on the assessment of the molecular coancestry-based coefficients. MolKin allows most variables to be computed, assigning the same weight to the information provided by each locus or weighting it according to its PIC (Botstein et al. 1980). Results obtained by weighting in accordance with the PIC are stored in the corresponding Access tables and .txt files under a name beginning with "W_." The user can also choose to compute molecular coancestry and genetic distances (and the corresponding standard errors) using bootstrapping, regardless of whether these values are PIC weighted or not. Results obtained by bootstrapping are stored in the corresponding Access tables and .txt files under a name ending in "Boot." The names of the Access tables and .txt files containing the results of the computations are usually self-informative regarding their content.

How to Use MolKin: Short Overview

MolKin starts by simply double-clicking on the program icon. After that, the initial screen of MolKin allows the user to find the .txt or .xls input data file in the corresponding directory and, for the .xls input files, to choose the worksheet on which the session will be carried out. After loading the input file, some descriptive computations of the input data are carried out and the user is presented with three different menus: Distances, Individual Kinship, and Rarefaction.

By clicking on the Distances menu, the user obtains, at an individual or population level, the matrices containing molecular coancestry coefficients, the kinship distance (D_k), Wright's (1978) F-statistics, Nei's minimum (D_m) and standard (D_s) distances (Nei 1987), the Reynold's distance (D_R), and the shared allele distance (D_{AS}) (Chakraborty and Jin 1993). By clicking on the Bootstrapping submenu, the user can obtain the majority of parameters using bootstrapping, without having to previously obtain the direct results. A detailed discussion on interpretation of the D_k and molecular coancestry matrices with respect of classical genetic distances can be found in Álvarez et al. (2005).

MolKin's second menu is the Individual Kinship menu. This has two submenus: the Between Individuals submenu, and the Mean Molecular Kinship submenu. The Between Individuals submenu is designed to help breeders in the management of a given population: the user selects any couple of individuals to be mated to obtain the corresponding molecular coancestry coefficients. To facilitate interpretation of the results, the average values for the whole analyzed population are presented on the screen. The Mean Molecular Kinship submenu computes the average molecular coancestry between each individual and all the others in the entire population or in the subpopulation in which the individual is classified (unweighted or PIC weighted). An example of how the average molecular coancestry can be used for conservation purposes is given in Table 1 for the rare Xalda sheep breed (Álvarez et al. 2004; Goyache et al. 2003). A total of 16 Xalda male individuals are candidates to be selected as parents for the following generation and genotyped with a set of 14 microsatellites (Álvarez et al. 2004, 2005). The 16 candidate individuals were analyzed jointly with 148 additional individuals representative of the live Xalda population. If conservation of genetic diversity is the breeding goal, individuals with the lowest average molecular coancestry values should be selected. Six of the 16 candidate individuals have average molecular coancestry values below the mean molecular coancestry of the whole genotyped population. Consequently these six individuals should be selected for reproduction.

MolKin's third menu is the Rarefaction Method menu, which gives descriptive statistics on genetic diversity (observed and expected heterozygosity, average PIC and average number of observed alleles per locus [rarefacted or not] for the analyzed populations). MolKin allows the user to fit a particular sample size (g) for rarefaction.

Table 1. Average molecular coancestry for 16 male Xalda sheep individuals computed both for the set of 16 individuals (tested subpopulation) and for the 16 individuals analyzed jointly with 148 additional individuals representative of the live Xalda population (whole population)

Individual	Tested subpopulation	Whole population
M537	0.300	0.286
M227	0.317	0.288
M608	0.378	0.319
M645	0.391	0.349
M15	0.423	0.355
M208	0.423	0.355
M609	0.423	0.355
M541	0.416	0.367
M25	0.417	0.374
M19	0.445	0.380
M22	0.445	0.380
M63	0.445	0.380
M156	0.422	0.398
M292	0.430	0.415
M330	0.475	0.436
M23	0.475	0.450

The lower the average molecular coancestry, the lower the representation of a genotype in the population. Average molecular coancestry for the whole dataset is 0.357 (whole population). The first six individuals have average molecular coancestry values less than 0.357 and should be selected for reproduction.

General Comments

MolKin has inherited some routines written for the program ENDOG (Gutiérrez and Goyache 2005), designed to analyze pedigree information. MolKin is written in Visual Basic and runs on Windows 95/98/2000/NT/XP. A setup menu guides the user when installing the program. The program, user's guide, and example file can be downloaded free of charge at http://www.ucm.es/info/prodanim/Molkin2.zip. MolKin has been tested on several datasets and results were checked for consistency with alternative software whenever possible. Classical genetic distances (D_m, D_s, D_R, and D_{AS}) and F-statistics have been tested using the programs GENEPOP (Raymond and Rousset 1995), Fstat (Goudet 1995), and Populations (Langella 1999), although MolKin computes these parameters (except D_{AS}) on molecular coancestry instead of on allelic frequencies. The authors would appreciate being informed of any detected bugs. Although the example file provided with the program includes a very small dataset, MolKin is capable of handling large data files such as that previously published by Hanotte et al. (2002), including more than 2000 individuals from 58 cattle populations genotyped for 15 autosomal microsatellite markers.

Acknowledgments

This article was partially funded by grants from INIA (nos. RZ03-011 and RZ2004-00007-C02). The authors wish to thank several researchers for testing the development versions of the program and Iván Fernández for his kind support and help.

References

Álvarez I, Gutiérrez JP, Royo LJ, Fernández I, Gómez E, Arranz JJ, and Goyache F, 2005. Testing the usefulness of the molecular coancestry information to assess genetic relationships on livestock using a set of Spanish sheep breeds. J Anim Sci 83:737–744.

Álvarez I, Royo LJ, Fernández I, Gutiérrez JP, Gómez E, and Goyache F, 2004. Genetic relationships and admixture among sheep breeds from northern Spain assessed using microsatellites. J Anim Sci 82:2246–252.

Bennewitz J and Meuwissen THE, 2005. A novel method for the estimation of the relative importance of breeds in order to conserve the total genetic variance. Genet Sel Evol 37:315–337.

Botstein D, White RL, Skolnick M, and Davis RW, 1980. Construction of a genetic linkage map in man using restriction fragment polymorphisms. Am J Hum Genet 32:314–331.

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

Chakraborty R and Jin L, 1993. Determination of relatedness between individuals using DNA fingerprinting. Hum Biol 65:875–895.

Eding H, Crooijmans RPMA, Groenen MAM, and Meuwissen THE, 2002. Assessing the contribution of breeds to genetic diversity in conservation schemes. Genet Sel Evol 34:613–633.

Eding H and Meuwissen THE, 2001. Marker-based estimates of between and within population kinships for the conservation of genetic diversity. J Anim Breed Genet 118:141–159.

Fabuel E, Barragan C, Silio L, Rodríguez MC, and Toro MA, 2004. Analysis of genetic diversity and conservation priorities in Iberian pigs based on microsatellite markers. Heredity 93:104–113.

Felsenstein J, 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791.

Goudet J, 1995. Fstat version 1.2: a computer program to calculate F-statistics. J Hered 86:485–486.

Goyache F, Gutiérrez JP, Fernández I, Gómez E, Álvarez I, Díez J, and Royo, LJ, 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.

Guo X and Elston RC, 1999. Linkage informative content of polymorphic genetic markers. Hum Hered 49:112–118.

Gutiérrez JP and Goyache F, 2005. A note on ENDOG: a computer program for analysing pedigree information. J Anim Breed Genet 122:172–176.

Hanotte O, Bradley DG, Ochieng JW, Verjee Y, Hill EW, and Rege JE, 2002. African pastoralism: genetic imprints of origins and migrations. Science 296:336–339.

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

Kumar S, Tamura K, Jakobsen IB, and Nei M, 2001. MEGA2: molecular evolutionary genetics analysis software. Bioinformatics 17:1244–1245.

Langella O, 1999. Populations 1.2.28 (12/5/2002): a population genetic software. CNRS UPR9034. Available at http://www.pge.cnrs-gif.fr/bioinfo/populations/index.php.

Malécot G, 1948. Les mathématiques de l'hérédité. Paris: Masson et Cie.

Nei M. 1987. Molecular evolutionary genetics. New York: Columbia University Press.

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

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

Takezaki N and Nei M, 1996. Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. Genetics 139:457–462.

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

Toro M, Barragan C, Ovilo C, Rodrigañez J, Rodríguez C, and Silió L, 2002. Estimation of coancestry in Iberian pigs using molecular markers. Conserv Genet 3:309–320.

Toro MA, Barragan C, and Ovilo C, 2003. Estimation of genetic variability of the founder population in a conservation scheme using microsatellites. Anim Genet 34:226–228.

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

Received April 6, 2005
Accepted August 25, 2005

Corresponding Editor: Sudhir Kumar