Made in United States of America
Reprinted from THE JOURNAL OF HEREDITY
Vol. 86, No. 3, May/June 1995

GENEPOP (Version 1.2): Population Genetics Software for Exact Tests and Ecumenicism

M. Raymond and F. Rousset

GENEPOP (version 1.2) is a population genetic software package for haploid or diploid data that is able to perform two major tasks. First, it computes exact tests or their unbiased estimation for Hardy-Weinberg equilibrium, population differentiation, and two-locus genotypic disequilibrium. Second, it converts the input GENEPOP file to formats used by other popular programs, like BIOSYS (Swofford and Selander 1981), DIPLOIDL (Weir 1990b), LINKDOS (Garnier-Gere and Dillmann 1992), and Slatkin's (1993) isolation-by-distance program, thereby allowing communication between them (ecumenicism).

Input File

GENEPOP requires an ASCII input file with very simple specifications (explained in detail in the GENEPOP documentation). All kinds of missing data can be handled. Each allele is coded by only two numbers, so that no more than 99 alleles can be considered. The number of populations or loci is not limiting for most options. After checking the input file, GENEPOP displays a general menu with the following options.

Option 1: Hardy-Weinberg Test

Hardy Weinberg (or HW) test is performed for each locus in each population. If there are four alleles or less, the exact HW test is performed, as described by Louis and Dempster (1987). If more than four alleles are present, an unbiased estimation of the exact HW probability is performed using the Markov chain method described by Guo and Thompson (1992). In both cases, GENEPOP provides the probability of error when rejecting H_0 (i.e., HW equilibrium) and, if the Markov chain method has been used, the standard error (SE) of the estimate. Other classical parameters are also automatically computed: expected genotypic proportions, allele frequencies, observed and expected numbers of homozygotes and heterozygotes, and so on. There is no limitation for the number of populations or loci.

Option 2: Genotypic Disequilibrium

Genotypic linkage disequilibria are described on, for example, page 102 of Weir (1990a). They are defined in terms of two-locus genotypic counts, so that it is not necessary to know the gametic composition of double heterozygotes (which in general is not available), and their absence can be tested by contingency table analysis. GENEPOP automatically creates all contingency tables corresponding to all possible pairs of loci in each population and analyzes them with a Markov chain method to estimate (without bias) the exact P value as described by Raymond and Rousset (in press).

This option requires a memory space determined by the number of loci and the number of individuals in the largest population. If these numbers are too large, this option will not work.

Options 3 and 5: Exact Test for Population Differentiation

For each locus, GENEPOP automatically builds a contingency table describing the allelic composition in each population and tests without bias the lack of genic differentiation, as described by Raymond and Rousset (in press). In the output, GENEPOP provides a contingency table for each locus, the estimate of the probability of error when rejecting H_0 (i.e., no allelic differentiation), and its standard error (SE). Option 5 corresponds to the possibility of testing automatically allelic differentiation for all pairs of populations for all loci.

Option 4: Private Allele Method for *Nm* Estimate

This option provides a multilocus estimate of the effective number of migrants (*Nm*) according to Slatkin (1985) and Slatkin and Barton (1989). Four estimates of *Nm* are provided, three using the regression lines published in Barton and Slatkin (1986), and a corrected estimate using the values from the closest regression line as described by Barton and Slatkin (1986).

Option 6: Conversion for *F* statistics

GENEPOP can convert the input file into the format required by the DIPLOIDL program. This program is written in FORTRAN and is derived from the listing published by Weir (1990b). It was first typed and modified by J. Goudet (University of Lausanne, Switzerland). The source program (DIPLOIDL.FOR) and an executable program (DIPLOIDL.EXE, limited to 40 populations, 10 loci, and 15 alleles per locus) as given by Goudet are provided with GENEPOP (with authorization). Among other things, this program computes F statistics according to Weir and Cockerham (1984) and builds a bootstrap confidence interval according to Weir (1990b).

Option 7: Conversion for Slatkin's Isolation-by-Distance Program

GENEPOP can convert the input file into the format required by the DIST program (Slatkin 1993) to detect isolation by distance. The program DIST.CPP, written in C++, is a slightly modified version of Slatkin's DIST.C. Also included is the information file (DIST.EMA) provided by Slatkin. The executable program (DIST.EXE) has been compiled for a maximum of 35 populations, 30 loci, and 40 alleles per locus.

Options 8 and 9: Conversion for BIOSYS

GENEPOP can convert the input file into the format required by BIOSYS (Swofford

and Selander 1981), either the letter or the number code.

Option 10: Conversion for ANOVA on Heterozygosity

This option converts the GENEPOP input file into the format required by an ANOVA of a variable indicating heterozygosity according to, for example, Weir (1990a, p. 120–124).

Option 11: Conversion for *D* statistics

GENEPOP can convert the input file into the format required by LINKDOS, a PASCAL program described by Garnier-Gere and Dillmann (1992) and based on Black and Krafsur (1985). This program performs pairwise linkage disequilibria analyses in subdivided populations and Ohta (1982) *D* statistics. The original source LINKDOS program, written in PASCAL (LINKDOS.PAS), and an executable file (LINKDOS.EXE, compiled for 40 populations, 20 loci, and 18 alleles per locus) are provided with the authorization of their authors.

General Comments

The various programs in GENEPOP have been thoroughly tested. Option 1 has been tested by comparing results with those of the EXACTP step in BIOSYS (Swofford and Selander 1981) for two allele cases, and with data published in Louis and Dempster (1987) and Guo and Thompson (1992) for more alleles. Options 2, 3, and 5 have been tested by comparing results with published data on contingency tables (e.g., Mehta and Patel 1983). Pseudorandom numbers needed for the various Markov chains are generated as described by Marsaglia et al. (1990).

GENEPOP runs on IBM PCs and compatibles, without any need for extended memory. A computer at least as fast as a PC with an Intel 486 processor is preferable (but not required) to obtain accurate estimates within a reasonable length of time. GENEPOP is not protected and is available by anonymous FTP at ftp.cefe.cnrs-mop.fr or upon e-mail request. All sources (written in QUICK-BASIC or in TURBO-PASCAL), executable programs, and a short manual are provided. An example of an input file is also distributed, plus all the output files generated with each option. The other programs (DIPLOIDL, DIST, and LINKDOS) distributed with GENEPOP are provided as given by their authors, and all questions concerning these programs should be addressed directly to them.

The test used in options 2, 3, and 5 corresponds to an unbiased estimation of the *P* value of Fisher's test on R×C contingency tables. This test can be applied to any contingency table, not just population genetic data, using the STRUC program (Raymond and Rousset, in press), which is also provided with GENEPOP.

From the Institut des Sciences de l'Evolution, URA CNRS 327, Laboratoire de Génétique et Environnement, Université de Montpellier II (CC 065), Place E. Bataillon, 34095 Montpellier cedex 05, France. We thank P. David, E. Imbert, and S. Samadi for their contributions during a "Stage annexe du DEA" in 1993; A. Becher, D. Bourguet, J. Britton-Davidian, J. Carlier, C. Chevillon, J. Dallas, P. David, P. Dias, B. Dodd, R. Eritja, A. Estoup, A. Failloux, S. Goodman, J. Goudet, P. Jarne, I. Olivieri, N. Pasteur, F. Renaud, M. Slatkin, F. Thomas, and two anonymous reviewers for useful comments, suggestions, or tests on the various states of GENEPOP until the present version. This is paper no. 94.090 of the Institut des Sciences de l'Evolution.

The Journal of Heredity 1995:86(3)

References

Barton NH and Slatkin M, 1986. A quasi-equilibrium theory of the distribution of rare alleles in a subdivided population. Heredity 56:409–415.

Black WC and Krafsur ES, 1985. A FORTRAN program for the calculation of two linkage disequilibrium coefficients. Theor Appl Genet 70:491–496.

Garnier-Gere P and Dillmann C, 1992. A computer program for testing pairwise linkage disequilibria in subdivided populations. J Hered 83:239.

Guo SW and Thompson EA, 1992. Performing the exact test of Hardy-Weinberg proportions for multiple alleles. Biometrics 48:361–372.

Louis EJ and Dempster ER, 1987. An exact test for Hardy-Weinberg and multiple alleles. Biometrics 43:805–811.

Marsaglia G, Zaman A, and Tsang WW, 1990. Toward a universal random number generator. Stat Prob Letters 8:35–39.

Mehta CR and Patel NR, 1983. A network algorithm for performing Fisher's exact test in r × c contingency tables. J Am Stat Assoc 78:427–434.

Ohta T, 1982. Linkage disequilibrium due to random genetic drift in finite subdivided populations. Proc Natl Acad Sci USA 79:1940–1944.

Raymond M and Rousset F, in press. An exact test for population differentiation. Evolution.

Slatkin M, 1985. Rare alleles as indicators of gene flow. Evolution 39:53–65.

Slatkin, M, 1993. Isolation by distance in equilibrium and non-equilibrium populations. Evolution 47:264–279.

Slatkin M and Barton NH, 1989. A comparison of three methods for estimating average levels of gene flow. Evolution 43:1349–1368.

Swofford DL and Selander RB, 1981. Biosys-1: a FORTRAN program for the comprehensive analysis for electrophoretic data in population genetics and systematics. J Hered 72:281–283.

Weir BS, 1990a. Genetic data analysis. Sunderland, Massachusetts: Sinauer Associates.

Weir BS, 1990b. Intraspecific differentiation. In: Molecular systematic (Hillis DM and Moritz C, eds). Sunderland, Massachusetts: Sinauer Associates; 373–410.

Weir BS and Cockerham CC, 1984. Estimating F-statistics for the analysis of population structure. Evolution 38:1358–1370.

Received March 28, 1994
Accepted October 4, 1994