Computer Note

EPISTACY: A SAS Program for Detecting Two-Locus Epistatic Interactions Using Genetic Marker Information

J. B. Holland

Epistatic interactions among genes can play an important role in their phenotypic expression, in genotypic variation in populations, and in response to natural and artificial selection. Detection and estimation of epistasis by traditional biometrical methods can be difficult, however.

Estimation of genotypic values at many loci has become feasible with the advent of molecular marker technologies. Information from molecular marker studies can provide a more direct method to estimate epistasis (Cheverud and Routman 1995). Edwards et al. (1987) tested for epistatic interactions among all two-locus pairs assayed in their study, but only 20 marker loci were used. Damerval et al. (1994) tested for epistasis among all pairs of 109 loci and reported that many important epistatic interactions were detected, even among loci that did not have significant main effects. Li et al. (1997) reported that epistatic interactions among quantitative trait loci (QTL) affecting grain yield components in rice were important, but most of these interactions would have remained undetected had not all possible (4465) pairs of 95 random markers been tested for epistasis. Holland et al. (1997) tested all pairs of 561 loci for epistatic interactions and reported important epistatic interactions among QTL, particularly between pairs of loci in which at least one locus had no significant main effect. Therefore, by restricting tests for epistasis to loci which have significant main effects, it is likely that some important epistatic interactions will not be detected. Unfortunately the number of possible two-way tests among n loci is $n(n - 1)/2$. Thus, with data on 561 loci, Holland et al. (1997) had to perform 157,080 tests to search for epistasis among all possible pairs. A computer program is needed to systematically and efficiently perform the large number of tests that are required in some cases.

EPISTACY is a program designed to be implemented in SAS (SAS Institute 1988) to perform all pairwise tests among any number of marker loci to detect epistatic interactions, to select those marker locus pairs that have detectable interactions at a chosen level of significance, and to print out genotypic means and interaction statistics associated with selected pairs. The output includes estimates of the error variance, the overall interaction variance, and the interaction partial R^2 for each selected pair. The partial R^2 statistic is computed as the interaction partial (type III) sum of squares divided by the total sum of squares, and refers to the amount of phenotypic variance explained by the epistatic interaction after accounting for the main effects of the two loci considered (Holland et al. 1997). Users are expected to have some experience with SAS in order to input their data in the correct format and to modify some aspects of the program to suit their requirements. The program uses two macros, one nested in the other, to create pairs of loci to be used as independent class variables in analyses of variance implemented by SAS Proc GLM (SAS Institute 1988). Each macro must be invoked $n - 1$ times to test all possible pairs of n loci. Thus users must write out $2(n - 1)$ macro invocations as part of their modifications to the program. Properly used, the program will not make redundant tests.

Two basic versions of the program have been written: one designed for use with recombinant inbred (RI) line populations and one for F_2 populations. The RI program has the option of eliminating heterozygotes from the tests, such that only additive-by-additive forms of epistasis will contribute to the epistatic interaction variance. The F_2 program allows for partitioning of the interaction variance into components due to additive-by-additive, additive-by-dominant, dominant-by-additive, and dominant-by-dominant forms of epistasis through the use of contrast statements in Proc GLM. The program can be used to analyze other generations or mating designs, as well, but in some cases interpretation of results becomes more difficult. For example, if genotypic data are taken from F_2 individuals and phenotypic data from self-fertilized progeny of the F_2s, the additive-by-additive portion of epistasis can be interpreted easily, but the forms of epistasis involving dominance are less easily interpreted because only half of the selfed progeny of a heterozygous individual will be heterozygous. In addition, if the F_2 program is used, the contrast statements will produce correct results only if all nine progeny classes exist in the data. Therefore the contrast statements will produce correct results only if codominant markers are used.

The fact that the number of pairwise tests for epistasis increases proportionally to the square of the number of loci considered causes not only the technical difficulty of how to execute many tests, but also the problem of experiment-wise error rates. The issue of experiment-wise error rates in molecular marker studies is already complex (Churchill and Doerge 1994). Bonferroni-style significance levels (Rawlings 1988) will tend to be far too conservative because of the dependencies among tests due to linkage. Permutation tests (Churchill and Doerge 1994) could be applied to epistatic analyses, but probably would be excessively computationally intensive to be practical. A liberal, but reasonable, significance level for testing all possible pairwise interactions could be calculated by dividing the comparisonwise error rate by $g(g - 1)/2$, where g is the number of linkage groups or chromo-

somes being studied. Users are encouraged to consider the issue of experimentwise error rates before implementing the program.

A different computer program, Epistat, also designed to analyze epistatic interactions among quantitative trait loci, was published recently (Chase et al. 1997). EPISTACY differs from the program developed by Chase et al. (1997) in that it is designed to work in the SAS system; EPISTACY uses linear models and least squares statistics rather than the maximum likelihood methods employed by Epistat; and EPISTACY has the capability to analyze data from F_2 individuals and F_2-derived lines as well as recombinant inbred lines and to test for dominant forms of epistasis, unlike Epistat which uses data only from homozygous loci.

The program was designed to run under PC-SAS, but it also works on mainframe or Macintosh versions of SAS, with minor modifications. There is a technical problem encountered running the program under Windows 95, but software is freely available from the SAS Institute to solve the problem, and instructions on obtaining the necessary software are included with the program. Copies of the program can be obtained from the author by sending a 3.5 inch diskette or an e-mail to the author. Hard copies of the program are available on request from the author as well. Detailed instructions, including examples of datasets and modified programs and outputs will be distributed along with the program. One example program analyzes data from 150 maize plants in the F_2 generation genotyped at 114 loci. This program required approximately 90 min of real time to run and accessed approximately 50 MB of hard drive memory when executed on a PC with 16 MB RAM and a 120 MHz Pentium processor running SAS under Windows 3.1. The second example program analyzes data from 84 oat RI lines genotyped at 252 loci and required approximately 7.5 h to run and accessed 246 MB of hard drive memory on the same system.

From the Department of Agronomy, Iowa State University, Ames, IA 50011. This is journal paper no. J-17267 of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa (project no. 3368). Data from the maize F_2 population used in the example program were kindly provided by D. Asmono and M. Lee, Iowa State University. Data from the oat RI population used in the example were kindly provided by W. Siripoonwiwat, L. S. O'Donoughue, D. Wesenberg, D. L. Hoffman, J. F. Barbosa-Neto, and M. E. Sorrells, Cornell University, DNA Landmarks, Inc., and USDA-ARS Small Grains Research Facility. The helpful comments of two anonymous reviewers were greatly appreciated.

References

Chase K, Adler FR, and Lark KG. 1997. Epistat: a computer program for identifying and testing interactions between pairs of quantitative trait loci. Theor Appl Genet 94:724–730.

Cheverud JM and Routman EJ, 1995. Epistasis and its contribution to genetic variance components. Genetics 139:1455–1461.

Churchill GA and Doerge RW, 1994. Empirical threshold values for quantitative trait mapping. Genetics 138:963–971.

Damerval C, Maurice A, Josse JM, and de Vienne D, 1994. Quantitative trait loci underlying gene product variation: a novel perspective for analyzing regulation of genome expression. Genetics 137:289–301.

Edwards MD, Stuber CW, and Wendel JF, 1987. Molecular-marker-facilitated investigations of quantitative trait loci in maize. I. Numbers, genomic distribution and types of gene action. Genetics 116:113–125.

Holland JB, Moser HS, O'Donoughue LS, and Lee M, 1997. QTLs and epistasis associated with vernalization responses in oat. Crop Sci 37:1309–1414.

Li Z, Pinson SRM, Park WD, Paterson AH, and Stansel JW. 1997. Epistasis for three grain yield components in rice (*Oryza sativa* L.). Genetics 145:453–465.

Rawlings JO, 1988. Applied regression analysis: a research tool. Pacific Grove, California: Wadsworth and Brooks/Cole.

SAS Institute, 1988. SAS/STAT™ user's guide, release 6.03 edition. Cary, North Carolina: SAS Institute Inc.

Received December 10, 1996
Accepted October 20, 1997

Corresponding Editor: Robert Angus