

EPISTACY: A SAS program for detecting two-locus epistatic interactions using genetic marker information.

Version 2.0 Help Guide

J.B. Holland  
USDA-ARS  
Department of Crop Science  
North Carolina University  
Raleigh, NC 27695-7620  
Phone: (919) 513-4198  
Fax: (919) 515-7959  
Email: james\_holland@iastate.edu

and

H. Ingle  
The Bayer Corporation  
Clayton, NC

Original version note by J.B. Holland published in Journal of Heredity 1998 Volume 89 pp374-375.

## Introduction

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 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 (4,465) pairs of 95 random markers had not 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 perform systematically and efficiently 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 a 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 to eliminate 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 form 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  $F_2$ 's, 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 comparison-wise error rate by  $g(g-1)/2$ , where  $g$  is the number of linkage groups or chromosomes being studied. Users are encouraged to consider the issue of experiment-wise error rates before implementing the program.

A different computer program, "Epistat", also designed to analyze epistatic interactions among quantitative trait loci, has been 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" that uses data only from homozygous loci.

The program was designed to run under PC-SAS, but also works on mainframe or Macintosh versions of SAS as well 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 3-1/2-inch diskette or an email to the author. Hard copies of the program are available on request from the author as well. Detailed instructions, including examples of data sets and modified programs and outputs will be distributed along with the program. One example program analyzes data from 150 maize  $F_2$  progeny genotyped at 114 loci. This program required approximately 90 minutes of real time to run and accessed approximately 50 MB of hard drive memory when executed on a PC with 16 MB RAM, 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 hours to run and accessed 246 MB of hard drive memory on the same system.

### **Acknowledgements**

Data from the maize F2 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'Donoghue, 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'Donoghue 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.* Wadsworth and Brooks/Cole, Pacific Grove, CA.

SAS Institute Inc., 1988. SAS/STAT™ User's Guide, Release 6.03 Edition. SAS Institute Inc. Cary, NC.

### **What's new in Version 2.0?**

Version 2.0 was modified to eliminate the need to write out each locus name in each of two macro invocations. Instead, the two macros that executed the analysis in the original program have been replaced with a single macro that will analyze all pairs of marker loci. The new program is more efficient as well, because most of the data set manipulations are conducted only on the “significant” results. The new version requires the user to have a list of all of the marker locus names. The new program will read this list and use it to generate locus pairs in the analysis macro. Finally, the new program is written to read in data from open Excel spreadsheets. This may be easier for many users, and if not, the old data input statements can be used to read from text files.

### **Known Bugs**

#### *Problems running EPISTACY with SAS version 8.0*

SAS version 8.0 introduced an “enhanced” editor that highlights SAS programs with colored fonts to ease debugging. Unfortunately, there is a bug with this editor: when code is embedded in a macro, the program returns to the regular editor rather than the enhanced editor after the macro loops. So, the program quits after one macro loop. SAS technical help says they are aware of the problem and will fix it in future versions of SAS. Until then, the problem is avoided by simply running the program from the regular editor. On the SAS menu, choose “View” then “Program Editor”, then open the program file and run it from that window.

#### *Problems Running EPISTACY Under Windows 95*

EPISTACY works optimally under Windows 3.1 and Windows98, but encounters a technical difficulty with memory allocation when run under Windows 95. The problem occurs with all SAS programs that use more than 500-600 macro loops, as do the example programs written for EPISTACY. The following information was provided by The SAS Institute:

SAS Note: D134

Title: SAS bug causes Windows 95 to hang in macros that loop enough times

Keywords: SYS.SYS BASE D0090627 DAV JAN97

V6-SYS.SYS-D134 UNOTE SHIP D134

6.11 WIN95 6.12 WIN95

BUG WINDOWS 95 HANG MACROS LOOP ENOUGH TIMES WNC CLOSE MEMORY LEAK  
INSUFFICIENT IML SCL

PRODUCT:BASE

DIVISION: PC SYSTEMS

PRIORITY: N/A

MODULE :SYS.SYS  
SHIP :Y

SUPPORT: CATES, M.  
STATUS: USAGE NOTE

BUGID: D0090627

Repeated invocations of the SAS System under Microsoft Windows 95 will result in a memory leak that will cause the system to malfunction. It has been determined that about 600 invocations of the SAS System in interactive mode or about 2000 invocations in non-interactive mode will reproduce this problem. In addition, any macro code which loops sufficient times (our tests show 500-600 would be enough) will also reproduce the problem.

SAS will bring up a blank dialog box with a close button on it. Pressing the close button will clear the dialog box and end the SAS session. No other applications will start up, you will get an "insufficient memory" error message. You will also not be able to restart the PC from the Start button, you will have to cold boot.

Large IML loops have also encountered this problem.

We have a fix that can be downloaded from the SAS Institute, Inc. web site. Location is: <http://www.sas.com/techsup/download/pc/fiberfix.exe> This is a self-contained explodable file. Put it in a subdirectory by itself and double click on it. Readme.txt file contains password.

The "fiberfix" program provided by The SAS Institute has solved the memory leak problems for all EPISTACY programs that I have attempted to run under Windows 95.

### *Problems Running EPISTACY as a Batch Job Under UNIX or Other Mainframe Systems*

Note that unless the command "option nosource nonotes" is used at the beginning of an EPISTACY program which is to be run as a batch job on a mainframe system, the "log" file produced by the program will become too large and cause the program to exit without successfully completing execution. The command 'dm "log; clear"' that is suggested for use under a Windows or Macintosh operating system will be ineffective at suppressing the log file produced by a batch program.

A problem with use of the "options nosource nonotes" command is that it eliminates all error messages that are normally printed to the log file. If there are errors in the program, this will make troubleshooting the program exceedingly difficult. It is recommended that users begin by running sample programs with only a small subset of all of the marker loci to be analyzed ultimately and allow the log file to be printed as usual. Users can also set the selection criterion for the probability level much higher than is to be used ultimately, so that the program output can be checked on a subset of the markers. Under these conditions, troubleshooting the program and the output is made much easier. Once the user is confident that the program is executing as desired, the options "nosource" and "nonotes" can be used at the beginning of the program to suppress the log file, the probability criterion can be set at the appropriate level, and the full set of marker loci can be included in the macro invocations to test all possible pairs of loci.



## Examples of Program Modifications and Sample Data Sets

### Example 1: F2Example.sas

#### Data Files

We have Version 2.0 set up to read data in from an open Excel spreadsheet. This requires a few changes in the infile statement that was traditionally used to read data from text files. Also, we assume that the spreadsheets are organized such that progeny or lines are listed in different rows of the first column, and genotypic and phenotypic data listed in following columns. This is different from the format used for the MAPMAKER/QTL program. The data for both example programs were originally set-up in the MAPMAKER format and had to be altered. The original file with the maize F<sub>2</sub> data, for example, appeared as:

```
data type F2 intercross
150 125 63
#Dwi Asmono's F2 and F2:3 data set, Mo17H99 crosses (file \...\mapnew5.raw).

*UMC94
HAHAABHHAHAHAHAHAHHHHBHHBHHANHAHHHHBHHBHAHAHHHHHAABBBHAAAAHAANHA
ABHHHAHHANHHABHHAHAHAHHHHHAHHHHBBAHHHAHHHHHHHHHHBHHBHAHHANHBVHHHHBHHHHHAHBAHHAHAH
HAHHBHB
*UMC164
HAHAABVHHAHAHAHBAHABHVBHHBHHBAHAAHHAHHHHBHHHHHAHAHHHHHAABVHHAHAHAHAHHH
ABHHHAHHANHHABHHAHAHAHHHHHHHAHHHHBBAHHHAHHHHHHHHHHBHHBHHANHBVHHHHBHHHHHAHBAHHAHAH
HAHHBHB
...etc...
*ISU149
CACCCACCCACAACCCCAACCCACCAAACCCACCCCCCCCCACCCCCCCCCCCCCACCCAACCCCC-
CCCCCCCCCCCCCACACACAAACCCCAAAACACCAACCCCCCCACCCCCCCCCCCCCACCACCCACACCACCCCCCCACC
ACAACCC
*PLHT          190    240    215    195    210    210    215    210    230    215    210
                225    225    240    220    215    245    240    225    225    175    215    225
                215    170    240    210    210    215    205    185    240    205    235    230
                200
...etc...
                235    235    220    245    210    210    240    210    190    215
```

To convert this into a form suitable for SAS, the data were imported into a spreadsheet program and each progeny was represented by a single column. This data matrix data was transposed in the spreadsheet program, and a new column containing the F<sub>2</sub> progeny identification numbers was added. Some of the marker loci in the original data (eg., ISU149) were dominant loci. The data from these dominant loci were eliminated from the data set because codominant loci are required to test for epistasis in the F<sub>2</sub> generation.

There were many columns of genotypic data, so the data were split into two parts, each containing

about half of the marker data on all progeny. The data are included as two different sheets, MAIZEF2A and MAIZEF2B, within the same Excel spreadsheet, “MaizeF2Data.xls” The data appear as:

PROG	UMC94	UMC164	UMC157	NPI234	UMC13	P1	NPI429	NPI236
1	HH	A	A	A	A	A	H	B
2	AA	A	H	H	H	H	H	A
3	HH	A	A	H	H	H	H	B

etc...

A third sheet was made in this same file (“MaizeF2Data.xls”), containing the names of all of the loci to be tested. It is called “MaizeLoci” and is simply a column listing all of the loci to be tested (order does not matter here):

UMC94  
 UMC164  
 UMC157  
 NPI234  
 UMC13  
 P1  
 NPI429  
 etc...

### *Program Modification*

“F2Example.sas” was written by modifying the more general “EPIF2.sas” program in the way outlined here. Since the data are not in a single file, two “infile” statements are used in the program “EPIF2.sas”. Because the genotypic data are in an alphabetical format, the “input” commands must declare the marker loci as alphabetic data, using the “\$” symbol. Also, since SAS does not accept variable names including the “.” symbol, the letter “z” was substituted for the “.” in the names of such loci in the “input” statements. For example, notice that “BNL6.32” was written as “BNL6z32” in the first input statement. The data from the two files are input into separate data sets, which are then merged into a single data set, named “all”. Also note that the directory specified in the “infile” commands should be altered to match the directory that the data files are in. The program modifications begin with the following commands to read in the list of locus names (“locname”) and to determine the total number of loci to be tested (“nummark”):

```
data loci;
FILENAME ALLLOCI DDE "EXCEL|[MaizeF2Data.xls]MaizeLoci!R1C1:R114C1";
INFILE ALLLOCI NOTAB DLM= '09'X DSD MISSEVER lrecl = 10240;
input locname $;

N=_n_;
call symput ('nummark', trim(left(N)));
```

```
run;

proc sort data=loci;
by locname;
run;
```

Next, the program creates global variable names for each locus that can be processed in the analysis macro:

```
data a;
set loci;
by locname;
if first.locname then do;
i+1;
put locname= ;
ii=left(put(i,3.));
call symput ('genmark' || ii, trim(uppercase(locname)));
end;
run;
```

Next, the first half of the data are read in from the sheet called "MAIZEF2A" in the file called "MaizeData.xls". Note that the file must be OPEN in excel for this method of inputting to work. For other data sets, one simply changes the designation of the rows and columns to read from. In this case, we read from Row 2, Column 1 to Row 151, Column 65, "R2C1:R151C65," of the designated sheet ("MaizeF2A") in the designated file ("MaizeData.xls"):

```
data one;
FILENAME one DDE "EXCEL|[MaizeData.xls]MaizeF2a!R2C1:R151C65";
INFILE one NOTAB DLM= '09'X DSD MISSEVER lrecl = 10240;
input PROG UMC94 $ UMC164 $ UMC157 $ NPI234 $ UMC13 $ P1 $ NPI429 $ NPI236 $
UMC37 $ PANXX2z3 $ ISU018 $ ISU019A $ ISU006 $ UMC86A $ BNL6z32 $ UMC53 $
UMC78 $ UMC131 $ UMC98A $ UMC4 $ BNL8z44B $ BNL8z15 $ UMC121 $ BNL8z35 $
UMC26 $ BNL3z18 $ Sh2 $ UMC123 $ PIO20713 $ BNL5z46 $ Bt2 $ NPI292 $ PIO10z25
$ UMC111 $ UMC86B $ BNL6z25 $ UMC72 $ UMC27 $ BNL10z06 $ Bt1 $ BNL10z12 $
ISU019B $ UMC51 $ UMC68 $ NPI235 $ UMC65 $ PL1 $ UMC21 $ NPI280 $ UMC62 $
AGP1 $ BNL15z40 $ UMC98B $ UMC110 $ BNL7z61 $ BNL8z39 $ BNL8z44A $ UMC35 $
BNL9z11 $ UMC103 $ BNL9z08 $ UMC48 $ NPI268 $ UMC7 $ ;
```

```
data two;
FILENAME two DDE "EXCEL|[MaizeData.xls]MaizeF2b!R2C1:R151C65";
INFILE two NOTAB DLM= '09'X DSD MISSEVER lrecl = 10240;
input PROG C1 $ BNL3z06 $ UMC153 $ BNL8z17 $ BNL14z28 $ NPI209 $ UMC64 $
ISU005 $ ISU012 $ NPI287 $ ISU049 $ ISU040 $ ISU064 $ ISU104 $ ISU141 $
ISU169 $ ISU98 $ ISU119 $ ISU150 $ ISU133 $ ISU115 $ ISU036 $ ISU069 $ ISU109
$ ISU120 $ ISU139 $ ISU116 $ ISU074 $ ISU152 $ ISU124 $ ISU136A $ ISU136B $
ISU032 $ ISU088 $ ISU045 $ ISU048 $ ISU075 $ ISU147 $ ISU138 $ ISU093 $ ISU058
$ ISU047 $ ISU046 $ ISU100 $ ISU053 $ ISU021 $ ISU132A $ ISU132B $ ISU168A $
```

UMC165 \$ PLHT;

**data all; merge one two; by prog;  
proc print;**

The “proc print” command should print out the combined data set to the “output” screen as follows:

```

                P                B
                A  I      B      N B  B  B
      U  U  N      N  N  N  I  S  I  U  N      U  U  L  N  U  N      N  U
      U  M  M  P  U  P  P  U  X  S  U  S  M  L  U  U  M  M  8  L  M  L  U  L  M
      P  M  C  C  I  M  I  I  M  X  U  0  U  C  6  M  M  C  C  U  Z  8  C  8  M  3  C
O  R  C  1  1  2  C  4  2  C  2  0  1  0  8  Z  C  C  1  9  M  4  Z  1  Z  C  Z  S  1
B  O  9  6  5  3  1  P  2  3  3  Z  1  9  0  6  3  5  7  3  8  C  4  1  2  3  2  1  H  2
S  G  4  4  7  4  3  1  9  6  7  3  8  A  6  A  2  3  8  1  A  4  B  5  1  5  6  8  2  3

1  1  H  H  A  A  A  A  H  B  B  B  B  B  B  H  H  H  H  H  H  B  B  B  B  B  H  A
2  2  A  A  A  H  H  H  H  A  A  A  A  A  A  A  A  A  H  H  H  H  H  H  A  A  H  H  H
3  3  H  H  A  A  H  H  H  B  B  H  A  A  A  A  B  A  A  H  H  H  H  H  A  A  A  H  A  A
4  4  A  A  A  B  H  H  H  B  B  B  B  H  H  H  H  A  H  B  H  H  H  H  H  H  H  B
5  5  A  A  A  A  A  A  A  A  B  B  B  B  B  H  B  H  A  A  A  H  H  H  H  A  A  H  B  A

... etc...

      I  I                I  I  I
      I  I  I  I  I  S  S  I  I  I  I  I  I  I  I  I  I  I  I  S  S  S  U
      S  S  S  S  S  S  U  U  S  S  S  S  S  S  S  S  S  S  S  S  S  U  U  U  M
      U  U  U  U  U  U  1  1  U  U  U  U  U  U  U  U  U  U  U  U  1  1  1  C  P
O  1  1  1  0  1  1  3  3  0  0  0  0  0  1  1  0  0  0  0  1  0  0  3  3  6  1  L
B  2  3  1  7  5  2  6  6  3  8  4  4  7  4  3  9  5  4  4  0  5  2  2  2  8  6  H
S  0  9  6  4  2  4  A  B  2  8  5  8  5  7  8  3  8  7  6  0  3  1  A  B  A  5  T

146 H  A  H  H  H  A  H  H  H  A  A  B  B  A  A  H  A  A  H  H  H  H  H  B  A  B  210
147 A  A  H  H  B  B  A  A  B  H  B  A  H  A  A  A  H  H  B  A  H  B  B  240
148 H  A  A  A  H  B  B  B  H  B  H  H  H  H  H  A  B  H  H  H  H  A  B  H  B  A  210
149 B  A  A  A  A  H  H  H  H  A  A  A  B  B  B  H  B  H  A  H  H  H  H  A  H  B  190
150 B  B  B  H  H  H  B  B  H  B  H  H  H  B  B  H  B  H  H  H  H  H  H  A  H  H  215

```

Following this is the macro (“f2”) that conducts the analysis and sorts the outputs. A few modifications to this macro may be necessary. In this example, we used the “A”, “B”, “H” notation for genotypic classes, whereas the original program assumes the notation 0, 1, 2 is used. This requires two changes within the “f2” macro.

First, the macro computes four contrasts for specific forms of two-locus epistasis: additive-by-additive (AxA), additive-by-dominant (AxD), dominant-by-additive (DxA), and dominant-by-dominant (DxD) interactions. In order to compute the contrasts correctly, one must know the order in which SAS lists the two-locus genotypes. The general program assumes that the genotypes were coded using the 0,1,2

notation, in which case the nine genotypic means would be listed by SAS in the order: (0-0), (0-1), (0-2), (1-0), (1-1), (1-2), (2-0), (2-1), (2-2). If allelic notation is used, this can be written as:  $A_{11}B_{11}$ ,  $A_{11}B_{12}$ ,  $A_{11}B_{22}$ ,  $A_{12}B_{11}$ ,  $A_{12}B_{12}$ ,  $A_{12}B_{22}$ ,  $A_{22}B_{11}$ ,  $A_{22}B_{12}$ ,  $A_{22}B_{22}$

In this example, however, the A,B,H notation was used, and SAS orders the genotypic class means as:  $A_{11}B_{11}$  (A-A),  $A_{11}B_{22}$  (A-B),  $A_{11}B_{12}$  (A-H),  $A_{22}B_{11}$  (B-A),  $A_{22}B_{22}$  (B-B),  $A_{22}B_{12}$  (B-H),  $A_{12}B_{11}$  (H-A),  $A_{12}B_{22}$  (H-B),  $A_{12}B_{12}$  (H-H). Therefore, the order of the coefficients in the contrast statement must be changed to reflect this:

```
contrast "AxA" &m1*&m2 1 -1 0 -1 1 0 0 0 0;
contrast "AxH" &m1*&m2 1 1 -2 -1 -1 2 0 0 0;
contrast "DxA" &m1*&m2 1 -1 0 1 -1 0 -2 2 0;
contrast "DxD" &m1*&m2 1 1 -2 1 1 -2 -2 -2 4;
```

Also note that it is a good idea to keep the zero's at the end of these contrast lists, even though they are not required by SAS to compute the contrast. The reason is that it is possible that all nine genotypes may not be represented for some locus pairs. In such cases, the order of the genotypic means may change, and the contrast coefficients will be incorrect. By keeping all of the zero's, SAS expects nine different means, and if all nine are not present, it will not compute the contrast. Having no contrast calculated is preferable to having incorrect contrasts computed.

Second, we need to change the coding for the two-locus genotype means in the macro:

```
data geno1; set means;if geno1 = "A" and geno2 = "A"; geno00 = lsmean;proc
sort; by locus1 locus2;
data geno2; set means;if geno1 = "A" and geno2 = "H"; geno01 = lsmean;proc
sort; by locus1 locus2;
data geno3; set means;if geno1 = "A" and geno2 = "B"; geno02 = lsmean;proc
sort; by locus1 locus2;
data geno4; set means;if geno1 = "H" and geno2 = "A"; geno10 = lsmean;proc
sort; by locus1 locus2;
data geno5; set means;if geno1 = "H" and geno2 = "H"; geno11 = lsmean;proc
sort; by locus1 locus2;
data geno6; set means;if geno1 = "H" and geno2 = "B"; geno12 = lsmean;proc
sort; by locus1 locus2;
data geno7; set means;if geno1 = "B" and geno2 = "A"; geno20 = lsmean;proc
sort; by locus1 locus2;
data geno8; set means;if geno1 = "B" and geno2 = "H"; geno21= lsmean;proc
sort; by locus1 locus2;
data geno9; set means;if geno1 = "B" and geno2 = "B"; geno22 = lsmean;proc
sort; by locus1 locus2;

data allgeno; merge geno1 geno2 geno3 geno4 geno5 geno6 geno7
geno8 geno9;by locus1 locus2;
rename _name_ = trait; drop stderr lsmean geno1 geno2;
```

To invoke the macro that will do the analyses, one must specify the trait (dependant variable) to be analyzed, and also the the threshold probability value (alpha) that is used to declare an interaction significant. The program deletes results from all locus pairs that are not significant at the end of each invocation of the “var1” macro. The user is responsible for determining an appropriate alpha level for their situation. In general, however, testing of more locus pairs should require a more conservative (lower) threshold probability value. In this example, we want to analyze the trait called “PLHT” and use a threshold p-value of P=0.001, so we invoke the “f2” macro with this command:

```
%f2(plht, 0.001);
```

### *Program Output*

The output from the program will include the results of the “proc print” statement on the complete data set, already discussed, and the following statistics:

OBS	LOCUS1	LOCUS2	TRAIT	DFERR	SSERR	DFINT	SSINT	FINT	PROBINT
1	BT1	BNL8Z39	PLHT	126	31795.58	4	5651.90	5.59936	.00035077
2	UMC78	UMC98A	PLHT	138	32620.18	4	5016.00	5.30506	.00052748

  

OBS	SSMOD	SSAXA	FAXA	PROBAXA	SSAXD	FAXD	PROBAXD	SSDXA
1	7384.79	122.093	0.48383	0.48797	8.21596	0.032558	0.85710	5355.67
2	10233.90	989.085	4.18433	0.04270	3.01418	0.012752	0.91026	446.60

  

OBS	FDXA	PROBDXA	SSDXD	FDXD	PROBDXD	SSTOTAL	PARTR2	GENO00
1	21.2235	0.00001	7.18	0.0285	0.86629	39180.37	0.14425	200.714
2	1.8894	0.17150	4752.49	20.1055	0.00002	42854.08	0.11705	223.947

  

OBS	GENO01	GENO02	GENO10	GENO11	GENO12	GENO20	GENO21	GENO22
1	218.667	230.000	220.217	218.690	214.167	195.714	217.000	235.000
2	226.000	191.667	225.227	212.935	218.750	208.333	223.125	203.684

This output indicates that only two locus pairs showed significant interaction at the threshold level, P=0.001. The first two columns indicate the loci involved in the interaction. The first interaction listed here is between “BT1” and “BNL8.39”. The trait “PLHT” (plant height) is indicated next. This is followed by the degrees of freedom for the error variance estimate, “DFERR”; the sum of squares for the error variance estimate, “SSERR”; the degrees of freedom for the interaction, “DFINT”; the sum of squares for the interaction, “SSINT”; the F-value for the interaction, “FINT”; and the probability value for the interaction, “PROBINT”. It is this last variable that determines whether or not a locus pair will

be retained in the output. Note that the “PROBINT” of both locus pairs in the output are less than 0.001. The output columns continue on another line, with the sum of squares for the full model, both locus main effects plus their interaction, “SSMOD”; the sum of squares for the additive by additive epistatic contrast, “SSAXA”; the F-value for the AxA contrast, “FAXA”; the P-value for the AxA contrast, “PROBAXA”; the sum of squares, F-value, and P-value for the additive by dominant contrast, “SSAXD”, “FAXD”, and “PROBAXD”; the sum of squares, F-value, and P-value for the dominant by additive contrast, “SSDXA”, “FDXA”, and “PROBDXA”; and the sum of squares, F-value, and P-value for the dominant by dominant contrast, “SSDXD”, “FDXD”, and “PROBDXD”. The total sums of squares, model plus error, “SSTOTAL” are listed next, followed by the partial R<sup>2</sup> value, which is the partial sums of squares for the interaction divided by the total sums of squares, “PARTR2”. Finally, the nine genotypic class means are listed: “GENO00”, corresponding to the genotype A-A, “GENO01”, corresponding to the genotype A-H, etc.

### *Interpretation of the Output*

Interpretation of the form of the epistatic interaction detected by the program can be aided by the results of the contrasts and by the genotypic means printed in the output. First, notice that although the four contrasts used are orthogonal if there are equal numbers of genotypes in each class, this condition is not expected to be met, and the contrasts are not, in fact, orthogonal for that reason. For example, the sum of squares for the first interaction, between “BT1” and “BNL8.39” was 5651.90. The sum of the four contrast sums of squares, however, was  $122.09 + 8.22 + 5355.67 + 7.18 = 5493.16$ .

The first interaction is significant because of dominant by additive epistasis. The DxA contrast was the only one significant for that locus pair, and its sum of squares is much greater than those for any other contrast. Dominant by additive epistasis implies that the expression of dominance at the first locus is affected by the homozygous genotype at the second locus. The dominance of the first locus can be estimated from the genotypic means printed to see how it was affected by the second locus. For those genotypes homozygous for the “A” allele at locus BNL8.39, dominance at the BT1 locus can be estimated as  $GENO10 - (GENO00 + GENO20)/2 = 220.217 - (200.714 + 195.714)/2 = 22.0$ . This estimate is in the overdominance range. Dominance at the BT1 locus can be estimated for those genotypes homozygous for the “B” allele at BNL8.39 as  $GENO12 - (GENO02 + GENO22)/2 = 214.167 - (230.000 + 235.000)/2 = -18.3$ . This estimate is in the underdominant range. Thus, the expression of dominance at the BT1 locus changes from strongly overdominant to strongly underdominant, depending upon the homozygous genotype at the BNL8.39 locus, which is the cause of the significant dominant by additive contrast.

The second interaction, between UMC78 and UMC98A, seems due in small part to additive by additive epistasis (PROBAXA = 0.04), and in large part to dominant by dominant epistasis, because the sum of squares due to dominant by dominant epistasis is by far the largest of the contrasts, and the probability value is 0.00002. Dominant by dominant epistasis implies that the expression of dominance at the first locus is affected by the heterozygote at the second locus. Dominance at UMC78 for

genotypes heterozygous at UMC98A can be estimated as:  $GENO11 - (GENO01 + GENO21)/2 = 212.935 - (226.000 + 223.125)/2 = -11.6$ . This can be compared to the average of dominance estimated at each of the homozygous genotypes at UMC98A:  $\{[GENO10 - (GENO00 + GENO20)/2] + [GENO12 - (GENO02 + GENO22)/2]\}/2 = \{[225.227 - (191.667 + 208.333)/2] + 218.750 - [(191.667 + 203.684)]/2 = (25.2 + 21.1)/2 = 23.2$ . Thus, UMC78 is associated with overdominance for plant height when UMC98A is homozygous for either allele, but it is associated with underdominance when UMC98A is heterozygous. This is the cause of the significant dominant by dominant contrast.

## Example 2: EPISTACY.OAT

### Data Files

The data for this example come from Siripoonwiwat, W., L.S. O'Donoghue, D.Wesenberg, D.L. Hoffman, J.F. Barbosa-Neto, and M.E. Sorrells. 1996. Chromosomal regions associated with quantitative traits in oat. *J. Quantitative Trait Loci 2*, Article 3.

<http://probe.nalusda.gov:8000/otherdocs/jqtl/>. The data can be downloaded from the World Wide Web by accessing the "Grain Genes" map data page:

"<http://wheat.pw.usda.gov:80/ggpages/newggmaps.html>"

After eliminating the header rows, the genotypic data appear as:

```
*avn3      33033111313013330033112011111333313331103331123113110030113133...
*BCD1049   31331331333333311133333313311133233331311131113313333312113311...
*BCD1108   113311313133133332213313133331131313131311131111311111111311...
...etc...
*WG719B   13111331131313133313111131113133333313311331333111331133113111...
```

The phenotypic data are in a separate file, and appear as:

```
^GPI93 64.09 67.14 59.99 63.41 57.03 64.01 63.94 62.77 69.99 66.36 64.15 ...
^GPI94 66.06 70.02 67.78 68.25 59.93 62.00 67.40 61.20 70.79 63.82 63.91 ...
^GPI95 76.76 78.48 75.36 77.58 73.12 74.18 72.02 74.30 73.69 75.94 73.18 ...
...etc...
^YDI95 2730.61 2748.96 2819.17 3111.37 3082.00 3120.99 3021.06 2865.12 ...
```

As in the previous example, the original data files were imported into a spreadsheet program where each RIL was represented by a single column. This matrix of genotypic data was transposed in the spreadsheet program and a new column containing the RIL identification numbers was added. The phenotypic data were also imported into a spreadsheet file.

Again, the data set was very large, so it was split into five sheets of marker data ("RFLPA," "RFLPB," "RFLPC," "RFLPD," and "RFLPE") plus one sheet of phenotypic data ("Traits") within the spreadsheet file called "RILExampleData.xls".

## Program Modification

“ExampleRI.sas” was written by modifying the more general ““EPIL.sas”” program in the way outlined here. Since the data are not in a single file, six “infile” statements are used in the program. The data from the six files are input into separate data sets, which are then combined into a single data set, named “all” using the “merge” command. Also note that the directory specified in the “infile” commands should be altered to match the directory that the data files are in. The program modifications begin with the following commands:

```
data rflpa; infile "C:\rflpa.dat" missover;
input ril avn3 BCD1049 BCD1108 BCD1117 BCD115A
BCD1160 BCD1184 BCD1186 BCD1230B BCD1235 BCD1237 BCD1250 BCD1261A
BCD1265 BCD127 BCD1270 BCD1280A BCD1307 BCD1338A BCD1380A BCD1405
BCD1407 BCD1532A BCD1532B BCD1555 BCD1580 BCD1643A BCD1660 BCD1695
BCD1716A BCD1716B BCD1734 BCD1779 BCD1797A BCD1797C BCD1802 BCD1823A
BCD1829B BCD1851C BCD1856 BCD1860 BCD1871 BCD1872A BCD1872B BCD1876
BCD1882B BCD1882C BCD1897A BCD1931 BCD1950A BCD1950B;
if ril = "." then delete; proc print;

data rflpb;infile "C:\rflpb.dat" missover;
input ril BCD1968B BCD1968C
BCD269 BCD327 BCD342A BCD421A BCD454 BCD808A BCD897 BCD907 BCD961
BCD978 CDO1081 CDO1090C CDO1091 CDO1092 CDO1168A CDO1174A CDO1192A
CDO1192B CDO1196 CDO1199A CDO122 CDO1238 CDO1242 CDO1246A CDO1281
CDO1313 CDO1319A CDO1319B CDO1321A CDO1326 CDO1328 CDO1340 CDO1342
CDO1345 CDO1358 CDO1378A CDO1378B CDO1380 CDO1388 CDO1396 CDO1402A
CDO1403A CDO1403B CDO1407B CDO1414B CDO1419 CDO1423B CDO1428A CDO1428B
CDO1430;if ril = "." then delete; proc print;

data rflpc;infile "C:\rflpc.dat" missover;
input ril CDO1433 CDO1435B CDO1435C CDO1436C CDO1437B CDO1445A CDO1445B
CDO1454 CDO1464 CDO1466 CDO1467A CDO1471 CDO1495 CDO1509A CDO1509C
CDO1510 CDO1523A CDO1523B CDO187 CDO189 CDO220 CDO278 CDO304 CDO309A
CDO346B CDO348B CDO370 CDO373 CDO393D CDO395A CDO405 CDO412A CDO414
CDO420B CDO457A CDO460A CDO460B CDO480 CDO482A CDO482C CDO482D
CDO484A CDO539B CDO54 CDO542 CDO549A CDO57A CDO57C CDO57D CDO580
CDO585B CDO58A CDO58B CDO58C CDO590A CDO590B CDO595;
if ril = "." then delete; proc print;

data rflpd;infile "C:\rflpd.dat" missover;
input ril CDO608A CDO618A CDO638 CDO665A CDO673A
CDO708A CDO708B CDO718B CDO770B CDO772 CDO780 CDO795A CDO795B CDO82
CDO942 CDO962A CDO962B ISU0563B ISU0582A ISU0582B ISU1146A ISU1163
ISU1247A ISU1254B ISU1372B ISU1450 ISU1463 ISU1543A ISU1651 ISU1736
ISU1755B ISU1774A ISU1874A ISU1900A ISU1900B ISU1958B ISU1961 ISU2013
ISU2124A ISU2192A ISU2287 KSUA1 OG49 PTA71A PTB17 PX5 R209A R221
SAD1 UMN101 UMN106;if ril = "." then delete; proc print;

data rflpe;infile "C:\rflpe.dat" missover;
```

```

input ril UMN107A UMN107B UMN109 UMN114 UMN128 UMN13
UMN133B UMN162 UMN202 UMN207A UMN214A UMN214B UMN23 UMN287 UMN339A
UMN339B UMN341A UMN361 UMN363A UMN364A UMN388 UMN407 UMN409 UMN41
UMN420 UMN433 UMN498A UMN498B UMN5004 UMN5047 UMN706B UMN815B UMN826
UMN856A WG110B WG110C WG466 WG605 WG645 WG719A WG719B;
if ril = "." then delete; proc print;

data allrflp; merge rflpa rflpb rflpc rflpd rflpe; by ril;

data traits; infile "c:\traits.dat" missover firstobs=2;
input ril hda92 hda93 hda94 hda95 hdi93 hdi94 hdi95;

data all; merge allrflp traits; by ril;
proc print;

```

The “proc print” command should print out the combined data set to the “output” screen as follows:

```

          B      B      B  B B      B B
      B B B B B B B C B B B C B   B C B C C C B B C C B
      C C C C C C C D C C C D C B C D C D D C C D D C
      D D D D D D D 1 D D D 1 D C D 1 D 1 1 D D 1 1 D
  A 1 1 1 1 1 1 1 2 1 1 1 2 1 D 1 2 1 3 3 1 1 5 5 1
O  R V 0 1 1 1 1 1 1 3 2 2 2 6 2 1 2 8 3 3 8 4 4 3 3 5
B  I N 4 0 1 5 6 8 8 0 3 3 5 1 6 2 7 0 0 8 0 0 0 2 2 5
S  L 3 9 8 7 A 0 4 6 B 5 7 0 A 5 7 0 A 7 A A 5 7 A B 5

1  1 3 3 1 1 1 1 1 3 1 1 1 3 3 3 3 3 1 3 3 1 3 1 1 3
2  2 3 1 1 1 3 3 1 1 3 1 1 3 3 1 3 1 1 3 1 3 3 1 3 1 3
3  3 0 3 3 1 1 3 3 3 1 3 3 1 1 3 1 1 2 3 3 1 1 3 3 1 3
4  4 3 3 3 1 1 1 3 1 1 1 3 3 3 1 1 3 3 1 1 1 1 1 1 3 3
5  5 3 1 1 3 1 1 3 1 1 1 1 1 3 1 1 1 1 1 3 1 0 3 1 1 1

```

...etc...

```

      U U U U U U   U
      M M M M M M U M W W           W W
      N N N N N N M N G G W W W G G   H   H   H   H   H   H   H
      4 4 5 5 7 8 N 8 1 1 G G G 7 7   D   D   D   D   D   D   D
O  9 9 0 0 0 1 8 5 1 1 4 6 6 1 1   A   A   A   A   I   I   I
B  8 8 0 4 6 5 2 6 0 0 6 0 4 9 9   9   9   9   9   9   9   9
S  A B 4 7 B B 6 A B C 6 5 5 A B   2   3   4   5   3   4   5

81 0 0 0 0 0 0 0 0 1 3 3 3 1 1 3 60.67 63.70 53.67 61.33 60.00 55.67 .
82 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 58.33 63.33 68.70 60.33 61.67 65.00 61.67
83 1 1 3 3 3 3 3 1 3 3 1 3 1 3 3 60.00 63.67 69.00 59.33 61.67 62.67 60.00
84 1 3 3 1 1 1 1 1 3 3 1 1 3 1 3 62.33 65.00 71.30 60.00 61.33 65.00 62.00

```

Note that the genotypic data in the original files were coded 0 = no data, 1 = homozygous for allele of parent 1, 2 = heterozygous, 3 = homozygous for allele of second parent, which is different from the 0,1,2 notation used by the general program. One way to change this to eliminate missing genotypes

from the analysis is transform the data into the notation used by the program for each pair of loci, as they are named by the program. Therefore, we use a macro called “transform”, to update the data set called “all” to fit the notation used by the program.

In addition, an important issue that must be considered when analyzing recombinant inbred data is what to do with heterozygous genotypes. Heterozygous genotypes in recombinant inbred line populations are not necessarily informative as to gene action. The reason is that phenotypic data are collected on progeny developed from self-fertilization for one or more generations from a recombinant inbred parent. While the plant that gave rise to the line may have been heterozygous at some loci, the level of heterozygosity in the progeny tested will be lower. Therefore, the progeny may not express dominance at any detectable level. Interpretation of epistatic contrasts involving dominance is difficult in this case. Therefore, such contrasts are not recommended for RIL populations .

Therefore, for the analysis of recombinant inbred lines, one has two options:

- 1) Analyze all of the data for the two markers, including heterozygotes. From this analysis, the interaction sums of squares may have some contribution due to interactions between homozygous genotypes at one locus and heterozygous genotypes (presumably heterogenous lines) at the second locus.
- 2) Eliminate all of the heterozygotes from the analysis, in which case the estimated interaction variance is solely composed of additive by additive epistasis due to interactions among homozygous genotypes at the loci involved.

This example implements the second option. The data set is updated so that heterozygous marker loci are replaced with missing data points. This is also accomplished using the “transform” macro. This macro can be used to make other kinds of data transformations as needed.

```
%macro transform(dataset);  
    data &dataset; set &dataset;  
    %do i = 1 %to &nummark;  
        if &&genmark&i = 0 then &&genmark&i = ".";  
        if &&genmark&i = 1 then &&genmark&i = 0;  
        if &&genmark&i = 2 then &&genmark&i = ".";  
        if &&genmark&i = 3 then &&genmark&i = 2;  
    %end;  
%mend transform;  
  
%transform(all);
```

Once these data transformations are accomplished, the “ri” macro can be invoked to perform the epistasis analysis. As with the “f2” macro, the user must define the trait to be analyzed and the desired threshold p-value. For this example, we analyze trait “hda92” with a p-value of P=0.001:

```
%ri(hda92, 0.001);
```

### Program Output

The output from the program will include the results of the “proc print” statement on the complete data set, already discussed, and the following statistics:

OBS	LOCUS1	LOCUS2	TRAIT	DFERR	SSERR	DFINT	SSINT	FINT	PROBINT
1	BCD1265	WG645	HDA92	73	1238.98	1	229.708	13.5343	.00044517
2	BCD1280A	CDO82	HDA92	69	990.47	1	210.872	14.6902	.00027647
3	BCD1280A	UMN815B	HDA92	56	866.93	1	190.550	12.3088	.00089688
4	BCD1338A	BCD1779	HDA92	72	940.42	1	228.284	17.4779	.00008073
5	BCD1338A	CDO1342	HDA92	71	949.39	1	239.307	17.8966	.00006855
6	BCD1338A	CDO1454	HDA92	65	829.36	1	319.539	25.0436	.00000453
7	BCD1532A	UMN13	HDA92	47	794.57	1	211.514	12.5113	.00092208
8	BCD1695	UMN420	HDA92	63	1115.46	1	245.292	13.8538	.00042358
9	BCD1802	UMN23	HDA92	66	1162.08	1	229.637	13.0422	.00058814
10	BCD1851C	CDO370	HDA92	69	1209.75	1	213.265	12.1639	.00085247
11	BCD1851C	UMN5047	HDA92	62	1061.90	1	316.019	18.4511	.00006251
12	BCD1851C	WG605	HDA92	74	1352.98	1	232.705	12.7275	.00063645
13	BCD1856	ISU1254B	HDA92	58	998.17	1	290.399	16.8739	.00012694
14	BCD1856	ISU1736	HDA92	59	1061.34	1	257.503	14.3146	.00036371
15	BCD1871	CDO1509A	HDA92	61	1071.79	1	242.319	13.7914	.00044483
16	BCD1882C	CDO370	HDA92	69	1126.90	1	296.463	18.1525	.00006331
17	BCD1931	CDO1509C	HDA92	72	1286.34	1	225.752	12.6360	.00067297
18	BCD1950A	ISU1254B	HDA92	62	1067.85	1	250.093	14.5205	.00032105
19	BCD1968B	UMN41	HDA92	59	1001.54	1	220.237	12.9740	.00064899
20	BCD269	CDO58A	HDA92	63	967.33	1	200.892	13.0837	.00059371
21	BCD327	CDO1199A	HDA92	64	1087.05	1	206.739	12.1718	.00088330
22	BCD897	UMN856A	HDA92	54	919.90	1	214.711	12.6039	.00080680
23	CDO1168A	UMN41	HDA92	62	1062.05	1	205.789	12.0136	.00096468
24	CDO1199A	UMN214A	HDA92	76	1321.53	1	249.476	14.3471	.00030209
25	CDO1199A	UMN23	HDA92	65	1153.94	1	250.739	14.1238	.00036872
26	CDO1238	R221	HDA92	58	928.40	1	199.201	12.4447	.00082723
27	CDO1328	ISU1254B	HDA92	62	1044.84	1	275.156	16.3275	.00014944
28	CDO1328	ISU1900A	HDA92	62	962.50	1	222.813	14.3526	.00034510

OBS	SSMOD	SSTOTAL	PARTR2	GENO00	GENO02	GENO20	GENO22
1	300.125	1539.10	0.14925	59.5330	61.1900	63.8335	58.5431
2	462.990	1453.46	0.14508	60.4167	59.0183	59.2381	64.8336
3	403.031	1269.96	0.15004	60.4236	58.7882	59.7100	65.5333
4	243.558	1183.97	0.19281	62.5089	58.5243	58.5554	61.6311
5	251.487	1200.87	0.19928	63.9700	59.0957	58.6316	61.2217
6	338.769	1168.13	0.27355	63.7857	58.7065	58.5361	62.2213
7	259.684	1054.26	0.20063	58.4158	64.2300	61.4163	58.9680
8	303.404	1418.87	0.17288	61.7772	59.9800	57.9113	63.7847
9	299.537	1461.61	0.15711	61.0165	59.5896	59.5333	65.9256

10	213.909	1423.65	0.14980	65.9340	59.5948	59.7060	61.8000
11	340.569	1402.46	0.22533	67.2783	59.5005	59.9867	62.5755
12	237.176	1590.16	0.14634	64.2340	59.4924	59.5660	62.2941
13	351.998	1350.17	0.21508	60.0227	62.7107	63.0544	57.0479
14	332.594	1393.93	0.18473	60.1193	62.1959	63.1100	57.0479
15	280.708	1352.50	0.17916	58.7594	64.4100	61.8937	59.7333
16	296.758	1423.65	0.20824	66.6660	59.3617	59.5596	62.2105
17	226.591	1512.93	0.14922	62.0132	58.8007	59.2313	63.1785
18	318.930	1386.78	0.18034	60.6000	62.3750	63.4660	57.4000
19	289.297	1290.84	0.17062	63.2805	59.8660	56.6667	61.1107
20	230.816	1198.14	0.16767	59.4318	61.3711	62.3679	57.3069
21	357.539	1444.59	0.14311	60.0413	59.4893	60.4926	68.0550
22	334.296	1254.20	0.17119	58.6844	64.6675	61.5882	59.2857
23	236.191	1298.24	0.15851	62.9330	59.5600	57.4078	61.5333
24	312.505	1634.03	0.15267	60.4164	60.8691	67.3333	59.2900
25	303.067	1457.01	0.17209	61.1964	59.6281	59.0283	65.8144
26	376.201	1304.60	0.15269	61.9989	71.6700	62.0850	58.9643
27	341.942	1386.78	0.19841	60.4211	62.7857	63.4994	57.6471
28	325.201	1287.70	0.17303	60.3830	61.8608	63.0695	57.0447

OBS	LOCUS1	LOCUS2	TRAIT	DFERR	SSERR	DFINT	SSINT	FINT	PROBINT
29	CDO1345	CDO348B	HDA92	72	1269.93	1	225.089	12.7616	.00063565
30	CDO1388	ISU1254B	HDA92	62	1001.90	1	315.092	19.4987	.00004108
31	CDO1388	ISU1736	HDA92	61	1107.26	1	220.225	12.1324	.00092284
32	CDO1388	ISU1900A	HDA92	62	972.41	1	211.223	13.4674	.00050670
33	CDO1407B	ISU1900B	HDA92	60	979.41	1	250.902	15.3706	.00022947
34	CDO1419	CDO708B	HDA92	67	964.41	1	172.232	11.9654	.00094696
35	CDO1433	UMN287	HDA92	61	808.89	1	218.508	16.4781	.00014243
36	CDO1433	UMN856A	HDA92	53	776.77	1	214.200	14.6152	.00034867
37	CDO348B	R221	HDA92	59	983.44	1	254.565	15.2723	.00024249
38	CDO370	UMN388	HDA92	56	948.17	1	254.565	15.0349	.00027966
39	CDO414	UMN409	HDA92	62	1053.46	1	204.459	12.0331	.00095623
40	CDO482D	UMN815B	HDA92	54	880.88	1	198.979	12.1979	.00096255
41	ISU1254B	UMN409	HDA92	61	1035.96	1	265.852	15.6541	.00020091
42	ISU1736	UMN409	HDA92	60	1092.88	1	222.691	12.2259	.00089345
43	R209A	UMN407	HDA92	61	1136.50	1	226.910	12.1790	.00090383
44	UMN23	UMN815B	HDA92	61	1036.38	1	229.378	13.5010	.00050467
45	UMN388	UMN5047	HDA92	63	1086.05	1	271.381	15.7424	.00018844
46	UMN388	WG605	HDA92	61	1140.42	1	227.808	12.1852	.00090133

OBS	SSMOD	SSTOTAL	PARTR2	GENO00	GENO02	GENO20	GENO22
29	233.019	1502.95	0.14976	58.9165	62.6660	62.7295	59.5340
30	384.885	1386.78	0.22721	60.2594	62.8220	63.4894	57.2919
31	315.429	1422.69	0.15480	60.5236	62.0183	63.2028	57.2893
32	315.292	1287.70	0.16403	60.5395	61.8608	63.0361	57.0447
33	262.972	1242.38	0.20195	61.7368	59.8192	58.8329	65.7143
34	236.407	1200.82	0.14343	60.8744	59.1452	59.3706	63.9757
35	555.452	1364.35	0.16016	59.7177	59.2447	60.2046	67.4545
36	444.101	1220.87	0.17545	60.4914	59.4255	60.1900	67.1518

37	342.238	1325.68	0.19203	63.9987	57.9474	60.2547	62.3617
38	263.903	1212.07	0.21003	67.7500	60.0450	59.4505	62.0471
39	335.795	1389.25	0.14717	61.1250	62.9600	62.3335	57.1250
40	376.765	1257.65	0.15822	60.1773	59.3846	58.9622	65.6667
41	344.094	1380.05	0.19264	60.5968	62.9012	62.9521	57.1113
42	322.923	1415.81	0.15729	60.8235	62.9781	62.2353	56.9050
43	286.452	1422.95	0.15946	60.2694	61.7675	63.3956	57.3331
44	380.891	1417.27	0.16185	60.6663	60.1176	59.0600	66.3340
45	299.112	1385.16	0.19592	67.0550	59.5243	60.2971	62.2775
46	239.205	1379.62	0.16512	64.8522	59.2941	60.0367	62.6108

The output from the RIL analysis is simpler because it includes no contrasts and only four genotypic means, corresponding to the homozygous two locus genotypes. More interactions were detected at the  $P=0.001$  threshold level in this experiment compared to the maize experiment. Recall that, in this case, 252 loci were tested in 31,626 combinations compared to the 6,441 pairs tested in the previous experiment. So, some of the increase in interactions detected may be due simply to performing more tests.

The output lists the two loci involved in each interaction, the trait considered, the degrees of freedom of the estimate of error variance “DFERR”, the sum of squares of the error estimate “SSERR”, and the degrees of freedom for the interaction, “DFINT”. Notice that the DFINT for all pairs is 1 in this case. This is because heterozygous genotypes were not considered in the model, and this allows only one degree of freedom for the interaction. In this case, the interaction statistics are equivalent to those of the single degree of freedom contrast for additive by additive epistasis discussed in the previous example. DFINT is followed in the output by the sum of squares, F-value, and P-value of the interaction, “SSINT”, “FINT”, and “PROBINT”. On the next line, the sum of squares due to the full model - locus main effects plus interaction- “SSMOD”, and the total sum of squares, “SSTOTAL”, are listed. This is followed by the partial  $R^2$  of the interaction (interaction partial sum of squares divided by the total sum of squares), “PARTR2”. Finally, the four homozygous two-locus genotypic means are listed, “GENO00”, “GENO02”, “GENO20”, “GENO22”.

### *Interpretation of the Output*

The interpretation of these additive by additive epistatic interactions is often simpler than interpretation of dominant forms of epistasis. For example, the interaction with the highest partial  $R^2$  value was the interaction of BCD1338A and CDO1454 (observation 6), with a partial  $R^2$  of 0.27. This value is interpreted to mean that the interaction alone accounted for 27% of the total genotypic variation after the main effects of BCD1338A and CDO1454 have been accounted for. Additive by additive epistasis implies that the additive effect at one locus is affected by the homozygous state of the second locus. The additive effect of BCD1338A in combination with the “0” genotype at CDO1454 can be estimated as  $(GENO00 - GENO20)/2 = (63.7857 - 58.5361) = 2.62$ . The additive effect of BCD1338A in combination with the “2” genotype at CDO1454 can be estimated as  $(GENO02 -$

$GENO22)/2 = (58.7065 - 62.2213)/2 = -1.76$ . Thus, the additive effect of BCD1338A is positive in combination with one homozygote at CDO1454 and negative in combination with the other.

### **Analyzing other traits with the ExampleRI.sas program**

This can be accomplished by simply changing the trait named in the “ri” macro invocation, as follows. The only rule is that the macro invocations be placed after the macro definition (the piece of code that starts with “%macro ri(trait, pvalue);”):

```
%ri(hda93, 0.001);  
%ri(hda94, 0.001);  
%ri(hda95, 0.001);  
%ri(hdi93, 0.001);  
%ri(hdi94, 0.001);  
%ri(hdi95, 0.001);
```