Installation procedures for gassoc:

After downloading the gassoc_x.y.z.tar.gz file to your site, you will need to:

tar -zxvf gassoc_x.y.z.tar.gz

The created directories are doc, src and testdata. Please read the gassoc.doc file in the doc directory.

****

GASSOC Version 1.06

****

01 June 2001

Daniel J. Schaid, Charley Rowland and David Tines

****

1.0 Introduction:

The program "gassoc" (for genetic association tests) is a program for computing general score tests to test for linkage (in the presence of linkage disequilibrium - LD) between multiallelic genetic markers and disease, when diseased subjects (cases) and their parents are sampled.

Three general score statistics are computed (GTDT, GDOM,GREC), with an option to compute their empirical simulated p-values. In addition, maximum likelihood estimates of relative risk parameters are computed for two situations. First, additive effects of alleles on the log relative
risk (multiplicative effects on relative risk) are computed for all alleles of each marker. Second, if specified by the user, genotype relative risks are computed. Because a marker may have a large number of alleles, and hence a large number of genotypes, we currently compute genotype relative risks only when one allele is specified as "high-risk", and all other alleles are grouped together. Hence, with a specified risk allele, genotype relative risks are computed for the homozygous and heterozygous genotypes with the high-risk allele. This is a new feature as of version 1.06.

It is critical to recognize that statistical tests for this type of data are sensitive only to the occurrence of both linkage disequilibrium and linkage. There are two types of "association" that can exist. The presence of linkage disequilibrium implies that the frequencies of particular haplotypes, composed of disease and marker

loci alleles, causes disease-marker allele associations between families (i.e., at the population level), whereas linkage causes disease-marker allele associations within families. Our reference to association implies both LD and linkage in this context.

2.0 History:

Changes and enhancements from version 1.0 to version 1.1:

The program has been updated to run the above analyses for a series of markers contained in a LINKAGE formatted input file (prior to processing by makeped).

When running the program using the command line options as input, a parameter file describing the input data file is required.

An interactive input mode has been added if the user wishes to be prompted for the various program options.

Sending the simulated score statistics to a file and/or specifying a non-zero beta vector for the simulations may only be done when using the interactive input mode and specifying only 1 marker to analyze.

The results may be sent to a user-specified output file when using the interactive input mode.

The allele with the highest frequency for a particular marker is chosen to be the baseline category for relative risk estimates.

The seeds for random number generation are printed on the output for ease of reproducing results at a later date.

Changes and enhancements from version 1.01 to version 1.02:

Fixed bug involving error message for maxiter in int_input.c
Fixed bug involving calculation of p-values in as239.c (the bug resulted in some of the p-values for the chi-square distribution to be off by a small amount (generally < .01).

Changes and enhancements from version 1.02 to version 1.03:

Changed format of p-values on output to show additional decimal places.

Changes and enhancements from version 1.03 to version 1.04:

Eliminated a restriction on the length pedigree and person ids. (previously ids were limited to 4 digit integers).
Added message stating how many cases were excluded due to not having parental information (this was previously counted in the "not used due to missing alleles" message).

Changes and enhancements from version 1.04 to 1.05:

Increased Maximum line length for pedigree file to 700.
Allow for tab as well as space delimeted columns in pedigree file

Changes and enhancements from version 1.05 to 1.06

Marker Labels will be printed on output when gassoc is initiated from the command line and the labels are specified in the parameter file.

Added the option to generate genotype relative risk estimates based on GEN coding of genotypes as described by Schaid DJ(1996). See updated documentation below for details on how to use this new feature.

2.0 Using gassoc:

Gassoc may be used either interactively or by command line options.

2.1 Using gassoc interactively:

Gassoc is invoked by typing "gassoc" <enter> at the command line. The following illustrates an interactive session with annotated explanations in square brackets.

```
Name of the input file>

  [This file should be in LINKAGE format with only one disease
locus (locus type 1) and an arbitrary number of marker loci
with numbered alleles (locus type 3). The file name can be up
  to 25 characters in length.]

Any liability classes? (Y/N)>

  [Specify whether the input data file contains a column
 containing the liability classes for the disease locus.]

Relative column position of loci:
 Disease locus>
 First marker locus>
 Last marker locus>

   [Specify the disease and marker loci to be analyzed
 in relation to each other (ignoring the first 5 columns of
 the file and the column containing liability classes if it
 exists)  For example, if one wanted to analyze all 5 markers
 in the "test.pre" data used as an example below, one would
 enter the following:
       Disease locus>  1
       First marker locus> 2
       Last marker locus> 6

   If one wanted to analyze only the first marker:
       Disease locus>  1
```

```
            First marker locus> 2
            Last marker locus> 2
         ].


      Perform GRR analysis? (Y/N)>

        Which marker for GRR analysis?>
          [if GRR analysis is requested, only the marker specified
             will be analyzed.  The number indicating the relative
             column position of the desired marker locus should be
entered
             here]

       Provide label for GRR marker>
            [Label to be printed on output]

       Which allele is of interest>
            [specify the numeric code for the "high-risk" allele of
interest.
           If "2" is the allele of interest then the genotype
           relative risks presented will be for 2X vs. XX, and 22
           vs. XX, where X stands for any allele other than allele "2"]


      # of simulations (0 if no simulations)>


      Seeds for random number generator (integers between 1 and 30,000):
      Enter seed 1>
      Enter seed 2>
      Enter seed 3>

         [These prompts are given only if simulations were specified
        above.]

      Output simulated score statistics to a file? (Y/N)>

         [This prompt given only if simulations are specified and if
        first marker to analyze=last marker to analyze (1 marker).]

      By default, the simulations test the null hypothesis: All betas=0.
      Simulate under an alternative hypothesis? (Y/N)>

         [This prompt given only if simulations are specified and if
        first marker to analyze=last marker to analyze (1 marker).]

      Print iteration history? (Y/N)>

      Change max. iterations for convergence(default=25)? (Y/N)>

      Redirect output from the screen to a file? (Y/N)>
```

2.2 Using gassoc from the command line:

When command line options are specified, the following two input files are required:

1: Input data file in LINKAGE format (but not processed by makeped).
2: Input parameter file in LINKAGE format.

The names of these two input files may be specified using the -inf and -par flags described below. Note that all markers in the input data file are evaluated, unless the -grrmarker and -grr options have been specified,in contrast to specifying a range of markers in the interactive mode.

Marker labels may be specified in the parameter file by placing the label immediately following a # sign on the line describing the marker locus. example: #Marker 1. When specifying a marker for GRR analysis on the command line, the label given on the command line must match the label in the parameter file. These labels are case sensitive.

Command line options:

-inf<infile>

specifies the file named by <infile> to be used as the input data file. Default=in.pre.

example : gassoc -infotherfile.pre

-par<parfile>

specifies the file named by <parfile> to be used as the input parameter file. Default=in.par.

-ith

prints the iteration history with the U vector,V matrix, parameter estimates and log likelihood at each iteration. See reference at end for definitions of U and V.

example : gassoc -ith

-maxiter#

specifies # of maximum iterations for convergence of maximum likelihood estimates. Default=25 iterations.

example : gassoc -maxiter50

Note that when -maxiter0 is chosen, the mle's and likelihood ratio statistics are not computed, but the score statistics will be computed.

example : gassoc -parotherfile.par

-sim#

specifies # of iterations for the simulation loop, and prints the simulated p-value for each of the GTDT,GDOM and GREC statistics. Default=0. Note simulated p-values are not computed for GRR analyses.

example : gassoc -sim1000

The random number generator requires three integer seeds between 1 and 30000. It is best to use larger values as seeds.

-seeda#

specifies # to be used as the value for first seed.

-seedb#

specifies # to be used as the value for second seed.

-seedc#

specifies # to be used as the value for third seed.

example :
gassoc -sim1000 -seeda14569 -seedb25653 -seedc19848

-grrmarker<marker label>
-grr<allele number>
[Specifying both of these options will produce genotype relative risk estimates in addition to the TDT analysis for the marker and allele specified. The marker label specified here must also be specified as a comment after the line describing this marker in the parameter file. Note that either both -grrmarker and -grr must be specified, or neither, or an error will occur]

Example -par file including marker labels:

```
6 0 0 5  << NO. OF LOCI, RISK LOCUS, SEXLINKED (IF 1) PROGRAM
0 0.0 0.0 0  << MUT LOCUS, MUT RATE, HAPLOTYPE FREQUENCIES (IF 1)
1 2 3 4 5 6
```

```
1   2  << AFFECTION, NO. OF ALLELES
 0.999900 0.000100   << GENE FREQUENCIES                  DC gene
 4 << NO. OF LIABILITY CLASSES
 0.0000 0.0500 1.0000
 0.0000 0.2000 1.0000
 0.0000 0.5000 1.0000
 0.0000 0.9000 1.0000 << PENETRANCES
3   4  #D6S299
 0.25 0.25 0.25 0.25  << GENE FREQUENCIES
3   4  #D6S276
 0.25 0.25 0.25 0.25  << GENE FREQUENCIES
3   4  #D6S105
 0.25 0.25 0.25 0.25  << GENE FREQUENCIES
3   4  #D6S273
 0.25 0.25 0.25 0.25  << GENE FREQUENCIES
3   4  #D6S291
 0.25 0.25 0.25 0.25  << GENE FREQUENCIES
0 0  << SEX DIFFERENCE, INTERFERENCE (IF 1 OR 2)
0.10 0.10 0.10  0.10 0.10 << RECOMBINATION VALUES
1 0.10000 0.45000 << REC VARIED, INCREMENT, FINISHING VALUE
```

2.2 Output:

When run with command line options, all output is sent to the standard output stream. Output can be directed to a file using the unix ">" command. When run interactively, the user is given the option of redirecting the output to a specified file.

By default, the program produces the following output:

summary information detailing the number of cases used and not used from the input file
allele label
parameter estimate (Beta)
relative risk [exp(Beta)]
standard error of the estimate [SE(Beta)]
standard normal Z value
2 sided p-value for Z

The allele with the highest frequency for a particular marker is chosen as the baseline category for relative risk estimates.
The program displays the likelihood ratio statistic, the covariance/correlation matrix of the parameter estimates, and score statistics with p-values for each of the GTDT, GDOM and GREC statistics.

Additional output is produced by the command line options -ith, -sim#, and
-grrmarker combined with -grr, or by using the interactive input method. This additional output is described below.

Additional options available with interactive input:

Note that these options are only available when analyzing only one

marker.

The user is given the option to specify that the score statistics from each iteration of the simulation be output to a specified file. This file will contain one line for each iteration of the simulation. Each line will contain 3 columns: GTDT, GDOM, and GREC score statistics respectively.

The user is given the option to specify a nonzero vector to be used as the beta vector in the simulation. After being prompted for the number of alleles for the marker (k), the user will then be prompted to enter the (k*(k+1)/2)-1 values for the vector.

NOTE: For K distinguishable alleles, the user needs to specify (K*(K+1)/2)-1 values for the vector. The first (K-1) betas are the main effects of alleles, and the remainder are betas for interactions of alleles. Default is a vector of zeroes (i.e., simulations under the null hypothesis of no associations).

Example: K=4 alleles, allele 1 is the reference allele, alleles 2 and 3 do not alter risk, allele 4 has twice the risk of allele 1 (ln(2)=.69), and there are no interactions. This requires 3 betas for main effects (0,0,.69) and 6 interactions (all set=0).

See Schaid DJ (1996), Table IIA for more complicated examples of the beta terms.

Help: At the command line type:

```
gassoc -h <enter>
```

This will display the following list of possible input options available with the program:

```
Usage:     gassoc

  [-ith]          :     print iteration history(default=No)
  [-maxiter#]     :     # = maximum iterations for convergence(default=25)
  [-inf<infile>] :     infile is input data file name (default=in.dat)
  [-par<parfile>]       :       parfile is input parameter file name
(default=in.par)
  [-sim#]         :     # = iterations for simulation loop (default=0)
  [-seeda#]       :     # = seed1 for random no. generator
  [-seedb#]       :     # = seed2 for random no. generator
  [-seedc#]       :     # = seed3 for random no. generator
  [-grrmarker<label>]   :   label for desired marker that matches the
marker
                        in the parameter file
```

[-grr<allele>] : "high-risk" allele of interest for grr analysis

3.0 Example Run of gassoc:
==========
% gassoc -inftest.pre -partest.par -sim1000 -seeda24689 -seedb18320 -seedc971

****
* gassoc Version 1.06
****

==========

ANALYSIS FOR MARKER: 2
==========

Summary Info:

# of valid lines in input file: 175

# of affected cases: 41
# of affected cases used in analysis: 22
# of affected cases not used: 19
not used due to missing parent or missing parent alleles: 19
not used due to case missing alleles: 0
not used due to inconsistent parent/case alleles: 0

Conditional Logistic:

Final estimates of Beta:

Allele	Beta	Rel. Risk exp(Beta)	SE(Beta)	Z	P(2-sided)
1	0.9510	2.5882	0.6911	1.3760	0.16882215
2	0.2661	1.3049	0.4704	0.5657	0.57162801
4	-0.6812	0.5060	0.7404	-0.9200	0.35757058

LR Statistic: 5.1662, df=3, p=0.160020733

Covariance/Correlation Matrix (*=Corr(Bi,Bj)):

0.4776	0.4524*	0.3724*
0.1471	0.2213	0.3085*
0.1906	0.1074	0.5482

Score Statistics:

	Score	df	P-value	Sim P-value(Simulations=1000)
GTDT:	4.8647	3	0.181975450	0.187000000
GDOM:	4.0036	3	0.261070514	0.244000000
GREC:	1.0769	3	0.782647661	1.000000000

Note: Seeds used for random# generation were 24689, 18320, 971

==========

ANALYSIS FOR MARKER: 3
==========

Summary Info:

# of valid lines in input file: 175

# of affected cases: 41
# of affected cases used in analysis: 20
# of affected cases not used: 21
not used due to missing parent or missing parent alleles: 21
not used due to case missing alleles: 0
not used due to inconsistent parent/case alleles: 0

Conditional Logistic:

Final estimates of Beta:

Allele	Beta	Rel. Risk exp(Beta)	SE(Beta)	Z	P(2-sided)
1	0.7687	2.1569	0.5345	1.4380	0.15043039
3	-0.0295	0.9709	0.6291	-0.0470	0.96254987
4	-0.3761	0.6866	0.7460	-0.5041	0.61417011

LR Statistic: 3.5523, df=3, p=0.314040722

Covariance/Correlation Matrix (*=Corr(Bi,Bj)):

0.2857	0.3224*	0.2706*
0.1084	0.3958	0.0873*
0.1079	0.0409	0.5565

Score Statistics:

	Score	df	P-value	Sim P-value(Simulations=1000)
GTDT:	3.4418	3	0.328393643	0.329000000
GDOM:	4.2683	3	0.233908850	0.234000000
GREC:	2.3333	3	0.506165219	0.818000000

Note: Seeds used for random# generation were 24689, 18320, 971

==========

ANALYSIS FOR MARKER: 4
==========

Summary Info:

# of valid lines in input file: 175

# of affected cases: 41
# of affected cases used in analysis: 21
# of affected cases not used: 20
not used due to missing parent or missing parent alleles: 19
not used due to case missing alleles: 1
not used due to inconsistent parent/case alleles: 0

Conditional Logistic:

Final estimates of Beta:

Allele	Beta	Rel. Risk exp(Beta)	SE(Beta)	Z	P(2-sided)
1	-0.4670	0.6269	0.5526	-0.8451	0.39807066
2	-0.2320	0.7929	0.5171	-0.4487	0.65365849
4	-0.3011	0.7400	0.6942	-0.4337	0.66451112

LR Statistic: 0.8457, df=3, p=0.838507034

Covariance/Correlation Matrix (*=Corr(Bi,Bj)):

0.3054	0.4497*	0.1132*
0.1285	0.2674	0.2518*
0.0434	0.0904	0.4819

Score Statistics:

	Score	df	P-value	Sim P-value(Simulations=1000)
GTDT:	0.8404	3	0.839778611	0.858000000
GDOM:	0.7687	3	0.856937723	0.866000000
GREC:	0.0526	1	0.818545808	1.000000000

Note: Seeds used for random# generation were 24689, 18320, 971

==========

ANALYSIS FOR MARKER: 5
======================

Summary Info:

   # of valid lines in input file: 175

   # of affected cases: 41
   # of affected cases used in analysis: 21
   # of affected cases not used: 20
      # not used due to missing parent or missing parent alleles: 20
      # not used due to case missing alleles: 0
      # not used due to inconsistent parent/case alleles: 0

Conditional Logistic:

  Final estimates of Beta:

Allele	Beta	Rel. Risk exp(Beta)	SE(Beta)	Z	P(2-sided)
1	-0.2971	0.7430	0.4634	-0.6411	0.52144373
3	0.2641	1.3022	0.6167	0.4282	0.66850948
4	1.3000	3.6692	1.2510	1.0392	0.29871923

  LR Statistic: 1.9631, df=3, p=0.580097445

  Covariance/Correlation Matrix (*=Corr(Bi,Bj)):
          0.2147   0.3485*   0.1332*
          0.0996   0.3804   0.3821*
          0.0772   0.2948   1.5649

Score Statistics:

	Score	df	P-value	Sim P-value(Simulations=1000)
GTDT:	1.9059	3	0.592158829	0.620000000
GDOM:	2.1263	3	0.546617574	0.548000000
GREC:	0.7778	2	0.677809578	1.000000000

Note: Seeds used for random# generation were 24689, 18320, 971

==========

ANALYSIS FOR MARKER: 6
======================

Summary Info:

   # of valid lines in input file: 175

```
   # of affected cases: 41
   # of affected cases used in analysis:  22
   # of affected cases not used: 19
      # not used due to missing parent or missing parent alleles: 19
      # not used due to case missing alleles: 0
      # not used due to inconsistent parent/case alleles: 0

Conditional Logistic:

  Final estimates of Beta:

                          Rel. Risk
    Allele   Beta          exp(Beta)     SE(Beta)      Z          P(2-
sided)
    2        0.6122        1.8445        0.9643        0.6349
0.52551412
    3        0.5086        1.6630        0.5765        0.8822
0.37764532
    4       -1.0821        0.3389        1.1707       -0.9243
0.35532037

  LR Statistic:             2.7540,  df=3,  p=0.431124442

  Covariance/Correlation Matrix (*=Corr(Bi,Bj)):
          0.9298            0.5745*           0.2373*
          0.3194            0.3324            0.2537*
          0.2679            0.1712            1.3704

Score Statistics:

                  Score      df   P-value          Sim P-
value(Simulations=1000)
  GTDT:           2.5893      3   0.459370996      0.476000000
  GDOM:           2.0868      3   0.554579757      0.568000000
  GREC:           0.3600      1   0.548506235      0.691000000

Note: Seeds used for random# generation were 24689, 18320, 971
```

3.1 Example run of gassoc to produce GRR statistics

```
% gassoc -inftest.pre -partest.par -grr2 -grrmarkerD6S105

***********************************************
*    gassoc Version 1.06
***********************************************


=====================================================================

ANALYSIS FOR Marker - D6S105, Locus =  4
=======================
```

```
Summary Info:
   # of validlines in input file: 175

   # of affected cases: 41
   # of affected cases used in analysis:  21
   # of affected cases not used: 20
      # not used due to missing parent or missing parent alleles: 19
      # not used due to case missing alleles: 1
      # not used due to inconsistent parent/case alleles: 0

GRR coding scheme

Conditional Logistic:

  Final estimates of Beta:

                          Rel. Risk
  Genetype   Beta          exp(Beta)     SE(Beta)      Z          P(2-
sided)
  2|X        0.0712        1.0738        0.5213        0.1367
0.89130272
  2|2       -0.1452        0.8649        1.0541       -0.1377
0.89047234

  LR Statistic:             0.0717,  df=2,  p=0.964779271

  Covariance/Correlation Matrix (*=Corr(Bi,Bj)):
          0.2717            0.4716*
          0.2591            1.1110

TDT coding scheme

Conditional Logistic:

  Final estimates of Beta:

                          Rel. Risk
    Allele   Beta          exp(Beta)     SE(Beta)      Z          P(2-
sided)
    1       -0.4670        0.6269        0.5526       -0.8451
0.39807066
    2       -0.2320        0.7929        0.5171       -0.4487
0.65365849
    4       -0.3011        0.7400        0.6942       -0.4337
0.66451112

  LR Statistic:             0.8457,  df=3,  p=0.838507034

  Covariance/Correlation Matrix (*=Corr(Bi,Bj)):
          0.3054            0.4497*           0.1132*
          0.1285            0.2674            0.2518*
          0.0434            0.0904            0.4819
```

Score Statistics:

	Score	df	P-value
GTDT:	0.8404	3	0.839778611
GDOM:	0.7687	3	0.856937723
GREC:	0.0526	1	0.818545808

4.0 Compiling gassoc:

The program gassoc has been compiled on a Sun Workstation with operating systems
SunOS 4.1.3, and an ANSII C compiler (acc), and Solaris v3.5 with the gcc compiler.
An accomanying makefile can be used to compile and link gassoc routines. Also, a dataset called test.pre and corresponding parameter file test.par are distributed, which should give the same results as those in the example above (except that simulated p-values may differ slightly due to the
simulation process).

5.0 References:

Schaid DJ (1996): General score tests for associations of genetic markers with disease
using cases and their parents. Genetic Epidemiology 13:423-449.

6.0 Bug Reports:

Comments to improve this program, including the reporting of bugs, can be sent to
schaid@mayo.edu.

PLEASE TELL US HOW WE ARE DOING!

We appreciate your interest in the software developed by the Mayo Statistical
Genetics Group, and would like to hear from you about how you plan to use this software, any errors that you find, and any suggestions for improvement.

Please fill out this form and mail it to rowland@mayo.edu.

<---------- CUT HERE ---------->

SOFTWARE PACKAGE: GASSOC

YOUR NAME:

INSTITUTION:

MAIL ADDRESS:

PHONE NUMBER:

FAX NUMBER:

E-MAIL ADDRESS:

BUG REPORT:

COMMENTS: