SASGENE: A SAS Computer Program for Genetic Analysis of Gene Segregation and Linkage

J. S. Liu, T. C. Wehner, and S. B. Donaghy

Gene segregation and linkage analysis is one of the fundamental methods in plant breeding and genetics research. When many pairs of gene loci in many progenies are involved in an experiment, the analysis can be time consuming. Two computer programs for gene linkage analysis have been previously published (Liu and Knapp 1990; Suiter et al. 1983; GMENDEL and LINKAGE-1, respectively). However, checking of the linkage phase (coupling, repulsion, or not useful) and eliminating useless recombination values were not handled in these programs. Useless recombination values include the case where both parents of the cross have dominant (or both have recessive) alleles or where the recurrent parent does not have a homozygous recessive condition for the two (or more) genes being studied. In addition, GMENDEL runs only on MS-DOS, requiring an Intel central processing unit, and LINKAGE-1 will not run on newer versions of some computer operating systems.

Since many researchers have access to computers that use the SAS (SAS Institute 1988) system, a SAS computer program called SASGENE was developed for genetic analysis of gene segregation and linkage. SAS programs run on virtually any computer operating system with only minor modifications by the user. Currently SAS runs on the following computer operating systems: Macintosh, DOS, Windows, UNIX, OS/2, MVS, CMS, OpenVMS, and VSE. SAS programs have the advantage that they can be updated easily by the scientist to run on new computer systems without requiring extensive work by a programmer. SASGENE is easily modified, such as adding a statement to print the intermediate results if more information were desired by the researcher, or providing analyses for linkage estimation of codominant genes and two gene goodness-of-fit tests. SASGENE is dimensioned for unlimited gene loci and unlimited individuals per generation and family, subject only to available disk space.

SASGENE consists of two SAS macros. The first macro, SGENE, performs single gene goodness-of-fit tests of the observed frequency data to the expected Mendelian segregation ratios, analyzing each trait involved in the experiment. The output of SGENE consists of the observed segregation numbers, chi square, and probability value by gene and family (Figure 1).

The second macro, LINKAGE, performs an analysis of gene linkage relationships. Chi-square and probability values are calculated to analyze gene pairs for independent assortment in the segregating generations (F_2, BC_{1P1}, and BC_{1P2}). Gene loci are checked in all possible pairs. Linkage was estimated using the chi-square method, a widely used standard for genetic data analysis (although it may produce inaccurate results in some cases). According to the phase (coupling, repulsion, or not useful), the recombination frequency (rf) and standard error (SE) are calculated using the following formulae (Sinnott and Dunn 1939; Weir 1994):

F_2 (repulsion):

$$\begin{aligned} rf &= p \\ &= \{-(bc + ad) \\ &\quad + [(bc + ad)^2 + ad(bc - ad)]^{1/2}\}^{1/2} \\ &\quad \div (bc - ad) \end{aligned}$$

F_2 (coupling):

$$rf = 1 - p$$

$$SE = \sqrt{(1 - p^2)(2 + p^2)/2n(1 + 2p^2)}$$

BC_1 (only coupling accepted):

$$rf = (b + c)/n$$

$$SE = \sqrt{rf(1 - rf)/n}$$

where a(A_B_), b(A_bb), c(aaB_), and d(aabb) are genotype segregation ratios in F_2 or BC_1. A sample output of linkage relation analysis for genes by LINKAGE is shown in Figure 2.

SASGENE requires an input data file that consists of plot number, replication number, plant number, family number, generation number, and gene (or trait) names. Families and generations can be assigned numbers, and genes are usually given letters corresponding to the trait categories. For example, the gene name might be "ll" for little leaf, and the categories "L" for little and "N" for normal leaf. Before analysis, observation data is converted to D (dominant) or R (recessive). An optional

Cucumber Gene Linkage Example
Single Locus Goodness of Fit Test
Probability >.05 is accepted as Single Locus
---------- GENE=SS FAMILY=44 ----------

GENERATION	N	DOMINANT	RECESSIVE	MISSING	EXPECTED	CHISQ	DF	PROB
P1	45	45	0	0				
P2	45	1	40	4				
F1	54	49	5	0				
F2	162	103	55	4	3:1	8.11	1	0.004*
BC1P1	81	78	3	0	1:0	0.11	1	0.73
BC1P2	81	38	42	1	1:1	0.20	1	0.65

Figure 1. Sample output of the single gene goodness-of-fit test by SASGENE part 1.

Cucumber Gene Linkage Example
Recombination Frequency (RF) Data in F2, & BC1 Population
Prob with * indicates gene pair might be linked

GENE1	GENE2	FAM	PHASE	N	A	B	C	D	MISS -ING	CHISQ	DF	PROB	RF	STD ERROR
						GENERATION=F2								
U	SS	30	C	162	69	27	24	36	6	75.8	3	0.000*	0.323	0.036
U	SS	44	C	162	77	27	26	28	4	35.5	3	0.000*	0.350	0.038
U	NS	28	C	162	89	16	18	21	16	24.3	3	0.000*	0.265	0.034
U	NS	30	C	162	74	22	33	27	6	35.0	3	0.000*	0.364	0.038
RC	NS	30	R	162	83	34	24	15	6	4.8	3	0.18	0.559	0.042

Figure 2. Sample output of linkage analysis by SASGENE part 2. A(A_B_), B(A_bb), C(aaB_), and D(aabb) are genotype segregation ratios in F_2 and BC_{1F1} or BC_{1F2}.

macro (convert.sas) is available to convert gene value to "D" or "R".

SASGENE is distributed free of charge. A sample dataset is available that illustrates use of the two macros. It also shows how to set up data properly for the analysis. The program files (readme.txt, startup.sas, convert.sas, sgene.sas, and linkage.sas) and sample dataset (example.dat) can be obtained from the World Wide Web at the following address: http://cuke.hort.ncsu.edu/wehner.html, or by sending a blank 3.5 inch floppy disk to T. C. Wehner with a note whether you prefer MS-DOS or Macintosh formatting.

From the Departments of Horticultural Science (Liu and Wehner) and Statistics (Donaghy), North Carolina State University, Raleigh, NC 27695. The research reported in this publication was funded by the North Carolina Agricultural Research Service. The use of trade names in this publication does not imply endorsement by the NCARS of the products named, nor criticism of similar ones not mentioned. Please address reprint requests to Dr. Wehner, Department of Horticultural Science, North Carolina State University, Raleigh, NC 27695-7609.

The Journal of Heredity 1997:88(3)

References

Liu B-H and Knapp SJ, 1990. GMENDEL: a program for Mendelian segregation and linkage analysis of individual or multiple progeny populations using log-likelihood ratios. J Hered 81:407.

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

Sinnott EW and Dunn LC, 1939. Principles of genetics. New York: McGraw-Hill.

Suiter KA, Wendel JF, and Case JS, 1983. LINKAGE-1: a PASCAL computer program for the detection and analysis of genetic linkage. J Hered 74:203–204.

Weir BS, 1994. Genetic data analysis: methods for discrete population data. Sunderland, Massachusetts: Sinauer.

Received December 4, 1995
Accepted September 23, 1996

Corresponding Editor: Robert Angus