A review of the 'Statistical Analysis for Genetic Epidemiology' (SAGE) software package

*Robert C. Elston and Courtney Gray-McGuire**

Department of Biostatistics and Epidemiology, Case Western Reserve University, Cleveland, OH 44106–7281, USA
* *Correspondence to*: Tel: +1 216 368 4314; Fax: +1 216 368 4880; E-mail: mcguire@darwin.epbi.cwru.edu

Date received (in revised form): 22nd July 2004

Abstract

The 'Statistical Analysis for Genetic Epidemiology' (S.A.G.E.) software package is an integrated, comprehensive package of computer programs designed to perform many of the different analyses required in the study of genetic epidemiology. It offers a graphical user interface for most platforms and, unlike many programs available in the public domain, is flexible in both receiving many types of input files and in allowing the user to choose among output files. All of the programs accept the same data files and together provide the means to perform familial correlation, segregation, linkage and association analyses, as well as many of the ancillary analyses that help achieve these goals. Many, but not all, of the same or similar analyses can be performed (with more difficulty) using publicly available freeware. The primary limitations of S.A.G.E. at present are the lack of software for estimating haplotypes or for identifying probable double recombinants in linkage analysis. S.A.G.E. is continually being extended and upgraded, however, with automatic downloading of the latest version always available to users.

Keywords: *familial correlations, segregation, linkage and association analyses, heritability, transmission disequilibrium test, allele frequency*

The 'Statistical Analysis for Genetic Epidemiology' (SAGE.)[1] software package Version 1.0 was introduced in 1987. Since then it has changed and developed, becoming almost unrecognisable, although its function has remained the same: to give genetic epidemiologists the tools they need for the analysis of family and pedigree data. Version 5.0 supports a full graphical user interface (GUI), with dialogue boxes and pull-down menus, on Windows, Digital Unix, Solaris and Linux platforms. Formatting of the data and the naming of variables (including marker loci and alleles) is virtually unrestricted. Reasonable default values of all options are indicated, but the user maintains wide flexibility in the analyses that can be performed. This is unavailable in any other software with similar functions. SAGE is continually being extended and upgraded, with automatic downloading of the latest version always available to users. There is an annual licence fee, which varies according to the number of analyses that can be simultaneously performed, but substantial academic discounts apply. Many (certainly not all) of its functions are available as freeware, but S.A.G.E. offers the advantage, otherwise unavailable to human geneticists and epidemiologists, of an integrated package of programs with a modern GUI and wide flexibility that all accept the same data files. The following is a list of programs currently available and a brief description of what each one does. This is followed by a partial description of how a 'FUNCTION' utility expands the capabilities of these various programs. Finally, the most commonly used freeware available, having similar (although not exactly the same) functions, will be listed.

- PEDINFO provides many useful descriptive statistics on pedigree data, including means, variances and histograms of family, sibship and pedigree sizes and counts of each type of relative pair. It identifies consanguineous matings, marriage loops and marriage rings. This allows the user quickly to describe the data that have undergone any particular analysis.
- FCOR calculates multivariate familial correlations with their asymptotic standard errors without assuming multivariate normality of the traits across family members.[2] It calculates familial correlations for all relative pair types available in the sample pedigrees. The covariance between any two correlation estimates is available, making it a simple matter to test whether any two correlations, obtained from family data, are significantly different. There is an option to output a file with a tabular structure for the correlations and their standard errors, making it easy to format results into tables for publication.
- SEGREG fits and tests Mendelian segregation models in the presence of residual familial correlations. The trait analysed can be continuous (for which regressive models[3] or the finite

polygenic mixed model[4] can be used), binary (for which a multivariate logistic model[5] or the finite polygenic mixed model[4] can be used) or a binary disease trait with variable age of onset (using the finite polygenic mixed model[4]). In this last case, it is possible to include in the likelihood information about the prevalence of the disease (even if that information is imprecise). This program can also be used for commingling analysis,[6] to predict the major genotype of any pedigree member and to automatically prepare penetrance files needed for model-based linkage analysis.

- MARKERINFO detects Mendelian inconsistencies of markers in pedigree data. By default, it assumes that all markers are codominant and error-free, but there is the flexibility to allow for markers that exhibit dominance or that are not error-free. The output is designed to help the user find the source of any inconsistency, even if it can only be detected by examining more than one nuclear family in the pedigree.
- FREQ estimates allele frequencies from marker data on related individuals with known pedigree structure[7] and, provided the markers are codominant, automatically generates marker locus description files used by GENIBD, MLOD and other S.A.G.E. programs. This is done dynamically, the program searching for all the different alleles that occur in the sample for each marker.
- GENIBD generates both single- and multi-point identity-by-descent (IBD) distributions for pairs of related individuals, using a variety of algorithms tuned for different types of pedigrees. Exact methods can be used for small pedigrees with loops,[8] and simulation methods are available for large extended pedigrees without loops. IBD sharing can also be interpolated between markers using either the Haldane or the Kosambi map function. The output also contains maternal and paternal IBD sharing values that can be used to assess parent-of-origin effects.
- RELTEST indicates how to reclassify pairs of relatives according to their true relationship using full multi-point genome scan data. The method is based on a Markov process model of IBD allele-sharing along chromosomes.[9] This program currently analyses four different types of putative pairs: full sib pairs, half sib pairs, parent offspring pairs and unrelated marital pairs. A summary file is produced that contains the pairs to be reclassified, together with their mean allele-sharing statistic, parent-offspring statistic and, for each individual, the percentage of marker data that is missing. This last feature enables the user to know whether the suggested reclassification should be made, because it can be unreliable if based on data from less than half the genome.
- SIBPAL is designed for the analysis of sib pairs, or larger sibships, to detect linkage. In the case of binary traits, mean and proportion tests are performed for affected pairs, unaffected pairs and discordant pairs, using probabilistic estimates of their allele sharing. In the case of quantitative traits (including binary traits as 0,1 variables), the various forms of Haseman–Elston regression[10–12] are available. Analyses can use either single- or multi-point IBD information, and models allow for multiple genetic loci—including epistatic interactions[13] and covariate effects. Asymptotic p values can be validated by obtaining p values from the appropriate permutation distribution.
- LODPAL performs analyses based on the lod score formulation for affected sib pairs.[14] The current implementation is of the general conditional logistic model,[15] including the one-parameter model that allows for the inclusion of all affected relative pairs, covariates[16] and epistatic interactions. There is also an option to include discordant and/or unaffected pairs in the analysis.
- LODLINK performs model-based lod score calculations for two-point linkage between a main trait and each of a set of markers. The main trait may be a marker or any other trait that exhibits Mendelian transmission. In the latter case, an output file from SEGREG, which includes trait allele frequencies and individual specific penetrance probabilities, can be used as input. LODLINK uses the genotype/phase elimination algorithms,[17,18] together with other enhancements, to perform linkage calculations. Maximised lods are converted to p values, both as upper bounds and based on asymptotic theory. Tests of sex and locus heterogeneity can be performed, the latter based on predefined groups of families[19] or using a mixture model.[20]
- MLOD performs exact multi-point model-based lod score linkage analysis on small pedigrees of arbitrary structure[8] and an approximate analysis on large pedigrees without loops using a Markov Chain Monte Carlo technique. Again, an output file from SEGREG can be used as input to describe the underlying trait locus inheritance model.
- ASSOC analyses the association between a continuous and/or binary trait and covariates, which can include marker phenotypes that have been transformed into quantitative covariates, from pedigree data in the presence of familial correlations. It performs likelihood ratio tests and obtains maximum likelihood estimates assuming, in the case of continuous traits, multivariate normality after either of two transformations (George–Elston[21] or Box–Cox[22]), whose parameters can be simultaneously estimated with all the other model parameters. These parameters include polygenic heritability and further familial correlations. Likelihoods can also be corrected for single ascertainment.
- TDTEX implements several asymptotic and exact versions of the transmission disequilibrium test (TDT)[23] for testing linkage between a marker and a disease locus in the presence of allelic association or linkage disequilibrium. The exact tests are useful in cases where few data are available or where there are many alleles at the marker locus. Different types of tests are available, including an exact test and a Monte Carlo randomisation test, as well as several exact and asymptotic marginal homogeneity tests.[24]

- AGEON fits an age-of-onset distribution[25] to sibship data comprising both affected and unaffected sibs, allowing for covariates that can influence the mean, variance or skewness of the onset distribution. It then calculates two new traits that can be used to achieve more power in Haseman–Elston regression linkage analysis: disease susceptibility allowing for age[26] and a measure of age of onset.[27]
- FUNCTION is an all-purpose utility that calculates new variables for analysis, eg trimmed, winsorised, mean and/or variance-adjusted variables (the adjustment being done separately for user-defined subclasses); quantitative variables defined on the basis of marker genotypes (dominant, additive and recessive allele indicators); and a transmitted allele indicator that allows ASSOC to perform pedigree TDT analysis.[28]

Some freeware is often used to perform many of the functions performed by SAGE PAP[29] can be used for segregation analysis, but is based on the usual mixed major gene/polygenic model, rather than on regressive, or finite polygenic mixed, models. It can also simulate phenotypes and estimate expected lod scores; PEDCHECK[30] and PREST[31] can be used to find Mendelian inconsistencies in a way similar to MARKERINFO, but without the detailed marker by marker and family by family output. RELCHECK[32] and RELPAIR[33] can be used to infer relationships within families, comparable to RELTEST, but consider more relationships and also consider pairs of persons across different families. Several linkage programs, including LINKAGE, MERLIN, GENEHUNTER, GENEHUNTER-PLUS, SOLAR, ALLEGRO, FASTLINK, VITESSE, SIMWALK2 and SUPERLINK, collectively perform analyses comparable to the SAGE linkage programs GENIBD, SIBPAL, LODPAL, LODLINK and MLOD, and have been reviewed recently.[34] The SAGE programs do not currently perform haplotype analysis or identify probable double recombinants, as do the SIMWALK2 and GENEHUNTER programs. FBAT[35]— and PBAT[36] and PDT[37]— perform association TDT-type analyses, respectively, on nuclear and extended pedigrees, comparable to the TDT kind of analysis ASSOC can perform when FUNCTION is used to generate transmitted allele indicators. FISHER[38] can calculate polygenic heritability under a slightly more stringent distributional assumption than that used by ASSOC. EDT[39] has some of the functions of TDTEX, but is based on logistic regression analysis using asymptotic results. Finally, GAP,[40] GAS[41] and ACT[42] are general program packages like S.A.G.E., but each is much more limited in scope. All of these programs, and many others, are listed on the Rockefeller University website.[43]

References

1. http://darwin.case.edu/sage.
2. Keen, K.J. and Elston, R.C. (2003), 'Robust asymptotic sampling theory for correlations in pedigrees', *Stat. Med.* Vol. 22, pp. 3229–3247.
3. Bonney, G.E. (1984), 'On the statistical determination of major gene mechanisms in continuous human traits: Regressive models', *Am. J. Med. Genet.* Vol. 18, pp. 731–749.
4. Fernando, R.L., Stricker, C. and Elston, R.C. (1984), 'The finite polygenic mixed model: An alternative formulation for the mixed model of inheritance', *Theor. Appl. Genet.* Vol. 88, pp. 573–580.
5. Karunaratne, P. and Elston, R.C. (1984), 'A multivariate logistic model (MLM) for analyzing binary family data', *Am. J. Med. Genet.* Vol. 76, pp. 428–437.
6. McLean, C.J., Morton, N.E., Elston, R.C. *et al.* (1976), 'Skewness in commingled distributions', *Biometrics* Vol. 32, pp. 695–699.
7. Boehnke, M. (1991), 'Allele frequency estimation from data on relatives', *Am. J. Hum. Genet.* Vol. 48, pp. 22–25.
8. Idury, R.M. and Elston, R.C. (1996), 'A faster and more general hidden Markov model algorithm for multipoint likelihood calculations', *Hum. Hered.* Vol. 47, pp. 197–202.
9. Olson, J.M. (1999), 'Relationship estimation by Markov-process models in a sib-pair linkage study', *Am. J. Hum. Genet.* Vol. 64, pp. 1464–1472.
10. Haseman, J.K. and Elston, R.C. (1972), 'The investigation of linkage between a quantitative trait and a marker locus', *Behav. Genet.* Vol. 2, pp. 3–19.
11. Elston, R.C., Buxbaum, S., Jacobs, K.B. and Olson, J.M. (2000), 'Haseman and Elston revisited', *Genet. Epidemiol.* Vol. 19, pp. 1–17.
12. Shete, S., Jacobs, K.B. and Elston, R.C. (2003), 'Adding further power to the Haseman and Elston method for detecting linkage in larger sibships: Weighting sums and differences', *Hum. Hered.* Vol. 55, pp. 79–85.
13. Tiwari, H.K. and Elston, R.C. (1997), 'Linkage of multilocus components of variance to polymorphic markers', *Ann. Hum. Genet.* Vol. 61, pp. 253–261.
14. Risch, N. (1990), 'Linkage strategies for genetically complex traits. I. Multilocus models', *Am. J. Hum. Genet.* Vol. 46, pp. 222–228.
15. Olson, J.M. (1999), 'A general conditional-logistic model for affected-relative-pair linkage studies', *Am. J. Hum. Genet.* Vol. 65, pp. 1760–1769.
16. Goddard, K.A., Witte, J.S., Suarez, B.K. *et al.* (2001), 'Model-free linkage analysis with covariates confirms linkage of prostate cancer to chromosomes 1 and 4', *Am. J. Hum. Genet.* Vol. 68, pp. 1197–1206.
17. Lange, K. and Boehnke, M. (1983), 'Extensions to pedigree analysis: V. Optimal calculation of Mendelian likelihoods', *Hum. Hered.* Vol. 33, pp. 291–301.
18. Goradia, T.M., Lange, K., Miller, P.L. *et al.* (1992), 'Fast computation of genetic likelihoods on human pedigree data', *Hum. Hered.* Vol. 42, pp. 42–62.
19. Morton, N.E. (1956), 'The detection and estimation of linkage between the genes for elliptocytosis and the Rh blood types', *Am. J. Hum. Genet.* Vol. 8, pp. 80–96.
20. Smith, C.A.B. (1963), 'Testing for heterogeneity of recombination fraction values in human genetics', *Ann. Hum. Genet.* Vol. 27, pp. 175–182.
21. George, V.T. and Elston, R.C. (1988), 'Generalized modulus power transformations', *Commun. Statist.* Vol. 17, pp. 2933–2952.
22. Box, G.E.P. and Cox, D.R. (1964), 'An analysis of transformations', *J. R. Stat. Soc.* Vol. 26, pp. 211–252.
23. Spielman, R.S., McGinnis, R.E. and Ewens, W.J. (1993), 'Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM)', *Am. J. Hum. Genet.* Vol. 52, pp. 506–516.
24. Bickeböller, H. and Clerget-Darpoux, F. (1995), 'Statistical properties of the allelic and genotypic transmission/disequilibrium test for multiallelic markers', *Genet. Epidemiol.* Vol. 12, pp. 865–870.
25. Pericak-Vance, M.A., Elston, R.C., Conneally, P.M. *et al.* (1983), 'Age-of-onset heterogeneity in Huntington disease families', *Am. J. Med. Genet.* Vol. 14, pp. 49–59.
26. Zhu, X., Olson, J.M., Schnell, A.H. and Elston, R.C. (1997), 'Genetic Analysis Workshop 10: Model-free age-of-onset methods applied to the linkage of bipolar disorder', *Genet. Epidemiol.* Vol. 14, pp. 711–716.
27. Hanson, R.L. and Knowler, W.C. (1998), 'Analytic strategies to detect linkage to a common disorder with genetically determined age of onset: Diabetes mellitus in Pima Indians', *Genet. Epidemiol.* Vol. 15, pp. 299–315.

28. George, V., Tiwari, H.K., Zhu, X. *et al*. (1999), 'A test of transmission/disequilibrium for quantitative traits in pedigree data, by multiple regression', *Am. J. Hum. Genet*. Vol. 65, pp. 236–246.
29. http://hasstedt.genetics.utah.edu/.
30. http://watson.hgen.pitt.edu/register/soft_doc.html.
31. http://utstat.toronto.edu/sun/Software/Prest.
32. http://www.biostat.jhsph.edu/~kbroman/software/.
33. http://www.sph.umich.edu/statgen/boehnke/relpair.html.
34. Dudbridge, F. (2003), 'A survey of current software for linkage analysis', *Hum. Genomics* Vol. 1, pp. 63–65.
35. http://www.biostat.harvard.edu/~fbat/fbat.htm.
36. http://www.biostat.harvard.edu/~clange/default.htm.
37. http://www.chg.duke.edu/software/pdt.html.
38. http://www.biomath.medsch.ucla.edu/faculty/klange/software.html.
39. http://www.mds.qmw.ac.uk/statgen/dcurtis/software.html.
40. http://icarus2.hsc.usc.edu/epicenter/gap.html.
41. http://users.ox.ac.uk/~ayoung/gas.html.
42. http://www.epigenetic.org/Linkage/act.html.
43. http://linkage.rockefeller.edu/soft/list.html.