

SEGLINK: A SAS® System MACRO for variance-components genetic linkage analysis

Jennifer H. Lin & Michael A. Province
Division of Biostatistics,
Washington University School of Medicine, St. Louis, MO

Abstract

The variance components approach to linkage analysis is one of the most powerful and robust methods for localizing genes for complex traits. Currently, there is not any particular SAS® System PROC able to perform these analyses. We have created a SAS® System MACRO, SEGLINK, which works interactively with two existing genetic packages, MAPMAKER/SIBS and SEGPATH. SEGLINK executes MAPMAKER/SIBS to get estimates of the probability of IBD (identity by descent) sharing for all possible sibpairs. These are used in turn along with the trait information by SEGPATH to conduct the variance-components linkage analysis. The MACRO uses SAS® dataset which contains individual phenotype/genotype information, along with two which contain the population gene frequency and the genetic map distance. The MACRO makes it practical and easy to conduct large scale, genome-wide scan for gene regions likely to influence any number of phenotypic traits and to conduct and manage these results as SAS® output datasets.

Introduction

The variance-components approach to detecting genetic linkage has become popular and more used in nonexperimental human genetic research. This approach has greater power than traditional methods in that it makes use of the information within a nuclear family or extended pedigree, without having to throw out individuals with partially missing values on some or entire measurement, nor requirement that the sibship size be the same in all families (Province, Rice, Borecki, Gu, & Rao, 1998; Province & Rao, 1995). In addition, variance-components approach makes no statistical assumption (e.g., polymorphic alleles, complete penetrance) for latent trait genes in comparison with traditional linkage analysis (Schork, 1993). Finally, variance-component approach is a true multi-point approach that utilizes a wide range of genome regions to model genetic influence of a complex disorder (e.g., heart disease and diabetes) by trait loci on specific chromosomal region(s) (Amos, 1994; Goldgar, 1990; Schork, 1993).

The purpose of this paper is to present a SAS® System MACRO (SEGLINK) that performs the analysis by working interactively with two genetic packages, MAPMAKER/SIBS (Kruglyak & Lander, 1995) and SEGPATH (Province, et al., 1998). This MACRO has the merit of dealing with complex dataset containing lists of markers on different chromosomes and multiple phenotypes, all of which are to be used for analysis in one run. As there is no SAS® application designed to perform the analysis mainly used for linkage studies, the MACRO can be a more useful tool which is easy to use and is efficient in automatic execution of several required steps.

The Statistical Model

The variance-components approach partitions the total variance (or phenotypic variance) into three major parts: (1) genetic variance due to major locus within the chromosomal region of interest; (2) genetic variance due to all other loci; and (3) variance due to random individual-specific effects (e.g., environment). A likelihood-estimated equation, based on a symmetric covariance matrix which defines the relationship for a nuclear family or pedigree i , can be expressed as:

$$\Sigma_i = (P_i h_g^2 + 2 F_i h_e^2) s^2 + I_i s_e^2, \quad (1)$$

where P_i is the matrix whose elements are the proportions of genes that pairs of relatives (e.g., siblings) share identity by decent (IBD) at the major locus; h_g^2 is the heritability due to the latent trait gene, g . F_i is the matrix of kinship coefficients; h_e^2 is the remaining heritability due to genetic influence other than the latent trait gene. s^2 is the observed phenotypic variance; s_e^2 is the random individual-specific effects; and I_i is the identity matrix.

Limitation of CALIS Procedure

The CALIS procedure, a widely used SAS/STAT® application in behavioral research, is capable of performing the covariance analysis and determine the unique contribution due to particular effect, as described above. However, the CALIS procedure has less practical use in genetic research because of

the limitation of data format. Unlike the genetic package such as SEGPATH, whose input dataset contains intact family structure with each individual being treated as an observation, the CALIS procedure considers one family as an observation (or one record per family). In that sense, SEGPATH easily deals with the circumstance when studied families are with unequal size, which may be difficult to handle under the CALIS procedure. Take one example for instance, given a study with N maximum numbers in families, the CALIS procedure builds up a $N(N-1)/2 \times N(N-1)/2$ symmetric covariance matrix, where each family provides with a $N(N-1)/2$ pairs of relatives. However, families with less than N members will not have that many pairs, which result in missing covariances for not given pairs. These families will end up be neglected under the CALIS procedure which omits observations with missing values (SAS/STAT® Version 6). Given only a handful of families that are available in genetic studies, the CALIS procedure may wind up with only very few observations (i.e., families with N family members) available for analysis, which loses not only statistical power, but also representative results.

Even replacing the covariances for missing pairs of relatives in families with certain hypothesized values cannot solve the problem. One way may be done is to replace with some close to 0 values (e.g., 10^{-7} or 10^{-9}) for missing pairs in any family. The procedure may bias the overall model by fixing all the missing covariances in different families to be the same. Besides, at the time when the replaced covariances are transformed to log likelihood estimates, it creates the issue as to determining a representative value for replacement, since close to 0 values can be widely ranged from 0 to $-\infty$ in log transformed scale. On the contrary, the SEGPATH package, used by the MACRO in the present study, avoids such disadvantages without having to duplicate missing pairs of family members during the analysis.

Method

To perform the variance-components approach using SEGPATH, users need to provide with the estimation of ibd probability for all possible relative pairs (P_i matrix), aside from other information such as phenotypic and environmental variances (see Equation 1). To do so, the SAS® System MACRO (SEGLINK) presented here first executes the other popular genetic package, MAPMAKER/SIBS, to get IBD estimates. Accompanied by other information, the IBD estimates are then fed into SEGPATH to compute unique variances for studied parameters. SEGLINK also reads results from SEGPATH and outputs a SAS® dataset for further use.

As seen from Table 1, four major SAS® datasets (four macro variables, **data**, **genemap**, **locdesc**, and **genefreq**) are requested to execute MapMaker/Sibs. The first input dataset (macro variable--**data**) contains individual genetic and phenotypic information (e.g., family id, personal id, father id, mother id, sex information) with each individual treated as an observation (or one record per individual). The other three datasets are related to genotype marker information. One of the them (macro variable--**locdesc**) contains marker names used in the study. The other (macro variable--**genemap**) provides map distance among genes on the same chromosome. In addition, the MACRO has the option of substituting no distance between any two adjacent genes (i.e., two genes are too close to be distanced greater than 0 cM) for a slightly greater than 0 cM distance (macro variable--**mindist**). The replacement with mindist value will make sure each of the two very close genes still gets separate estimation of ibd probability. Finally, a genetic frequency dataset (macro variable--**genefreq**) contains allele frequency for each marker. If in case there are missing gene frequencies (since gene frequencies may be obtained from other population), the MACRO provides with the option (macro variable--**rarepcnt**) to fill in the missing gene frequencies.

Table 1: Macro variables in the datasets:

DATA=	Input Family Data File.
FAMID=	Family (Pedigree) ID variable name.
ID=	Individual ID variable name.
FID=	Father's ID variable name.
MID=	Mother's ID variable name.
SEX=	Sex variable name.
MALE=	Code for Male for SEX=variable.
FEMALE=	Code for Female for SEX=variable.
PHENOS=	List of Phenotype Variables (quantitative).
MARKERS=	List of particular Genotype Marker Variables .
LOCDESC=	Marker Description file (input).
GENEMAP=	Genetic Map dataset (input) depending upon the source.
MINDIST=	Constant to use as the "minimum map distance" .
GENEFREQ=	Gene Frequency dataset (input);
RAREPCNT=	Constant Percentage to use if allele found which NOT represented in GENEFREQ= dataset.

SEGLINK first reads in these datasets and prepare two input files (macro variables--**sibped** and **sibloc**)

to be used by MAPMAKER/SIBS. By specifying either single- or multi-point approach for ibd estimation (macro variable--**point**), the MACRO calls and executes MAPMAKER/SIBS outside SAS® environment, and outputs ibd estimate file and summary file (macro variables--**sibibd** and **sibout**). Table 2 provides with the macro variables that are used when running MAPMAKER/SIBS.

Table 2: Macro variables used when running MAPMAKER/SIBS:

POINT= MULTI or SINGLE.

SIBSPED= Temporary ped file name for MAPMAKER/SIBS.

SIBSLOC= Temporary loc file name for MAPMAKER/SIBS.

SIBSIBD= Temporary ibd file name for MAPMAKER/SIBS.

SIBSOUT= Temporary stdout file name for MAPMAKER/SIBS.

Then, the MACRO reads in the generated output of ibd estimate produced by MAPMAKER/SIBS, and merges back with the original data (macro variable--**data**) to prepare for a complete dataset containing both ibd estimate as well as phenotypic information. The newest merged dataset (macro variable--**segdat**) with SEGPATH required format, and accompanied by a job file (macro variable--**segjob**), with description of to-be-estimated parameters, is used by the MACRO which then calls and executes SEGPATH, outside of the SAS® environment. The execution results in a list of output files, including result file and summary files (macro variable--**segsrt**, **segter**, **segcsv**, **segplx**, and **segout**). The MACRO reads in the result file and creates a SAS® output dataset (macro variable--**out**). The MACRO also has the option to plot the marker scores on each chromosome.

Table 3. Macro variables used when running SEGPATH:

SEGJOB= Input Job file name for SEGPATH.
SEGDAT= Input Datafile name for SEGPATH.

SEGTJF= Temporary Job file name for SEGPATH.
SEGSRT= Temporary Sorted Datafile name for SEGPATH.

SEGTER= Temporary Terse Summary Output file for SEGPATH.

SEGCsv= Temporary CSV Summary Output file for SEGPATH.

SEGPLX= Temporary Detailed Prolix Output file for SEGPATH.

(Cont. Table 3)

SEGOUT= Temporary stdout file for SEGPATH.

FMT= Output SAS format to output phenos/pis in SEG DAT= file

OUT= Output SAS dataset containing results.

Moreover, the MACRO provides the option of ascertainment correction (macro variables--**selvar** and **selvalue**) if users are interested in particular families (e.g., families with offsprings severely affected by disease) or individuals (see Table 4). Users only need to specify the phenotypic variable and the cutoff point for that variable. The MACRO will pick up the families or individuals who are of interest.

Finally, the MACRO also takes care of some other issues. For instance, the MACRO accommodates the different strategies used by the two genetic packages. MAPMAKER/SIBS always omits families with less than two offsprings' genotypes since these families are unable to produce IBD estimates. However, the MACRO has the option of adding these omitted families and feeds into SEGPATH (macro variables--**missgeno** and **nopairs**) (see Table 4). This option may contribute statistical power in estimating parameters other than the linkage ones, such as phenotypic means and variances.

Table 4. Other useful macro variables.

SELVAR= Var name denoting Ascertained value for this obs.

SELVALUE= Value of SELVAR= variable which indicates a ascertained person. All other values are random.

MISSGENO= DELETE or KEEP Individuals with phenotypes but missing Genotypes.

NOPAIRS= DELETE or KEEP Pedigrees with fewer than 2 sibs.

Conclusion

The MACRO provides with a simple and user-friendly way to conduct variance-components approach for linkage analysis. Users only need to prepare for four major SAS® datasets and let the MACRO performs all the procedures in one run, which results in a SAS® output or plots for marker scores, if required. It also shows the flexibility of dealing with complex data structure (e.g., unequal family size, list of markers and phenotypes). In sum, the MACRO not only saves users substantial time in working all the procedures step by step. It also helps users, who

may not have a lot of familiarities with the two genetic packages, still be able to accomplish the analysis and fit their own desired model.

ACKNOWLEDGMENTS

This paper was partially supported by NHLBI grant, HL56567 and NIGMS grant, GM28719.

SAS, SAS/STAT are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Contact Jennifer H. Lin, address: Box 8067, 660 S. Euclid Ave., St Louis, MO 63110. E-mail: jennifer@wubios.wustl.edu.

References

- Amos, C. I. (1994). Multivariate oligogenic linkage analysis of quantitative traits in general pedigrees. *American Journal of Human Genetics*, 54, 535-543.
- Goldgar D. E. (1990). Multipoint analysis of human quantitative genetic variations. *American Journal of Human Genetics*, 47, 957-967.
- Kruglyak, L. & Lander, E. (1995). Complete multipoint sib-pair analysis of qualitative and quantitative traits. *American Journal of Human Genetics*, 57, 439-454.
- Province, M. A., & Rao, D. C. (1995). A general purpose model and a computer program for combined segregation and path analysis (SEGPATH): Automatically creating computer programs from symbolic language model specifications. *Genetic Epidemiology*, 12, 203-221.
- Province, M. A., Rice, T, Borecki, I. B., Gu C., Rao, D. C. (1998). Multivariate and multipoint variance-components approach involving structural relationships for assessing quantitative trait linkage using SEGPATH. Paper submitted for publication.
- Sas Institute Inc. (1990), SAS/STAT User's Guide, Version 6, Fourth Edition, Cary, NC: SAS Institute Inc.
- Schork N. J. (1993). Extended multipoint identity-by-descent analysis of human quantitative traits: Efficiency, power and modeling considerations. *American Journal of Human Genetics*, 53, 1306-1319.