Paper 240-2009

Beyond SAS/GENETICS™

Amina Barhdadi, Yassamin Feroz zada, Marie-Pierre Dubé
Montreal Heart Institute and Université de Montréal, Montreal, Canada

ABSTRACT

The field of statistical genetics has developed considerably over the past five years due to the substantial advances in genomic technology. The volume and the nature of these data create a need for reliable computer software to perform statistical analyses. SAS/GENETICS offers a collection of procedures specifically tailored for genetic data such as ALLELE, HAPLOTYPE, CASECONTROL and FAMILY procedures.

This paper highlights the basic features of some of these procedures as we routinely apply them to the analysis of human genetic association studies. We show examples where SAS/GENETICS can be complemented with other SAS/STAT procedures such as LOGISTIC and PHREG to conduct analyses commonly used in human genetic studies. We also compare output results of SAS/GENETICS procedures to those from analogous analyses with SAS/STAT procedures.

INTRODUCTION

The goal of genetic association studies is to test for an association between one or more genetic marker and a phenotype of interest and to estimate the magnitude of the association. The genetic information at a marker in one individual consists of pairs of discrete measures, called alleles. We find two alleles per person belonging to each of the chromosome inherited from a person's two parents, except for the X chromosome in men who have a single X accompanied with a Y chromosome. Each pair of alleles in an individual is called a genotype, and a set of alleles at different markers present on a single parental chromosome is called a haplotype. The most commonly used markers are single nucleotide polymorphisms (SNPs) usually presenting only two possible alleles at the population level.

SAS/GENETICS provides a framework and the tools necessary to analyze pairs of alleles at markers and their relationship with a dichotomous trait (e.g. disease status). The ALLELE and HAPLOTYPE procedures are concerned with the analysis of genetic marker data whereas CASECONTROL and FAMILY procedures are used to study the relationship between genetic markers and trait. Moreover, the HAPLOTYPE procedure can also test for the association between a binary trait and multiple markers. All these procedures provide output that can be sent to the PSMOOTH procedure to adjust p-values for large numbers of tests conducted on multiple markers obtained in a genome scan.

This paper will be most useful to SAS users involved in genetic association studies. After a brief review of the basics of SAS/GENETICS procedures and a description of situations where those may be insufficient, we provide the syntax of solutions to these situations using other SAS/STAT procedures.

APPLICATIONS

1. MARKER PROPERTIES AND DATA QUALITY CONTROL

Proc ALLELE and proc HAPLOTYPE are used to calculate characterizing features of markers before testing for marker-disease associations. In addition to providing measures of marker informativeness, estimates of allele, genotype and haplotype frequencies, these two procedures perform tests for Hardy-Weinberg equilibrium (HWE) at each marker and linkage disequilibrium (LD) between pairs of markers.

Prior to analysis, genotype data undergoes a quality control assessment that can be achieved using outputs from Proc ALLELE and Proc HAPLOTYPE. Measures of marker informativeness provided by Proc ALLELE (PIC, Het and div) give an indication of the amount of heterozygosity in the sample at each marker. Markers with higher values of these measures are more informative and most valuable for association and linkage studies. Haplotype frequencies either observed or estimated (using EM algorithm) provided by proc HAPLOTYPE are used for examining linkage disequilibrium between markers. We provide here the SAS code used for the assignment of minor and major alleles at each marker.

```
/* Data one should contain the following variables: Snpname, individualid, allele1 and
allele2; and should be sorted by snpname and individualid. Missing alleles should be
written as ' ' */

/* Proc allele to get alleles frequencies*/
ods output allelefreq=allelefreq(rename=(locus=snpname));
proc ALLELE data=one tall marker=snpname indiv=individualid ;
var allele1 allele2;
run;
ods output close;

proc sort data=allelefreq ;
by snpname freq;
run;

/*Step 1: Find the major allele*/
data majorallele(keep=snpname allele freq rename=(allele=majorallele freq=majorfreq));
set allelefreq;
by snpname freq;
if last.snpname;
run;

/*Step 2: Find the minor allele*/
data minorallele(keep=snpname allele freq rename=(allele=minorallele freq=minorfreq));
set allelefreq;
by snpname freq;
if first.snpname;
run;

/*Step 3: Merge databases one, majorallele and minorallele*/
data two;
merge one(in=a) majorallele(in=b) minorallele(in=c);
by snpname;
run;
```

Additional data quality control assessments that cannot be achieved using outputs from ALLELE and HAPLOTYPE procedures are required in genetic association studies. In particular, duplication and relatedness between individuals are to be avoided in case-control studies and should be checked prior to the analysis. This can be done by computing the average identity by state (IBS) at each marker between all pairs of individuals. Average IBS for a pair of individuals is calculated as the sum of the number of identical-by-state alleles at each marker divided by twice the number of markers. Individuals with identity value above 99% are likely identical twins or duplicates and those with values above 86% are likely closely related. These IBS relationships can be converted to distances by subtracting them from 1 and the matrix of pairwise IBS distances are used as input for multi dimensional scaling (MDS) where the first two components of each subject are plotted. Outliers from the main cluster will be removed as ethnic outliers. The SAS code is provided below.

```
/*Data one should contain the following variables: SNPname, individualid (subject
identifier), allele1 allele2, and idnumber (a variable that orders individuals from 1
to the total number of individual). This dataset should be sorted by SNPname*/.

/*Macro to calculate the average IBS between each pair of individuals*/
%macro ibs;
%let nindiv=&nindiv;
%do i=1 %to &nindiv /*nindiv is the total number of individuals */;
data two(where=(idnumber=%eval(&i)) rename=(id=indiv allele1=a1 allele2=a2));
set one;
run;

data three;
merge two(in=a) one;
by SNPname;
if a;
if a1=allele1 and a2 ne allele2 then ibs=1;
if a2=allele2 and a1 ne allele1 then ibs=1;
if a1=allele1 and a2=allele2 and a1 ne a2 then ibs=2;
```

```
if a1 ne allele1 and a2 ne allele2 and a1 ne allele2 and a2 ne allele1 then ibs=0;
if a1=a2=allele1=allele2 then ibs=2;
if a1=' ' or a2=' ' or allele1=' ' or allele2=' ' then ibs=999 /*missing value*/;
run;
proc SORT;
by id;
run;

data four (keep=indiv ibs idnumber id avg_ibs );
set three (where=(ibs ne 999));
%let nSNP=&nSNP;/*number of SNPs*/
retain sum 0;
by id;
if first.id then sum=ibs; else sum=sum+ibs;
avg_ibs=sum/(2*&nSNP); /*average IBS*/
if last.id;
run;

proc APPEND base=ibs data=four;
run;
%end;
%mend ibs;
%ibs;
/*Multidimensional scaling */
data mds;
set ibs;
ibs=1-avg_ibs;
run;
proc TRANSPOSE data=mds out=matrix(drop=indiv _name_);
var ibs;
by indiv;
id id;
run;

proc MDS data=matrix condition=row level=ordinal coef=diagonal dimension=2 pfinal
out=out outres=res ;
run;
 /*Graphic*/
   ods html;
 %plotit(data=out(where=(_type_='CONFIG')),
           plotvars=dim2 dim1,  symbols="*",
           datatype=mds, place=0, labelvar=subject, ls=100,
           vechead=, plotopts=hzero vzero);
   run;
ods html close;
```

2. TESTING FOR MARKER-TRAIT ASSOCIATION

In genetic association studies, traits can be measured as continuous variables, as discrete numerical categories or as affected /unaffected indicator variables (binary trait). Analysis of continuous traits can be done using standard SAS/STAT procedures such as Proc GLM and Proc MIXED. Proc CASECONTROL and proc FAMILY take simple dichotomous indicators of disease status and use standard algorithms to compute statistics of association between these indicators and the genetic markers.

2.1. Description of Proc CASECONTROL

The establishment of an association between a disease and a genetic marker is commonly achieved through the use of case-control designs. Proc CASECONTROL in SAS/GENETICS provides different case-control tests between a genetic marker and a binary trait. Contingency tables are created with cases and controls representing the two rows and genotypes or allele categories for the columns. Tests for differences in these frequencies are performed using the usual Pearson chi-square. Another type of test provided by CASECONTROL procedure is Armitage's trend test for additive allele effects.

2.2. Proc CASECONTROL compared to Proc LOGISTIC

We compared results from the output of the Proc CASECONTROL genotype and trend test to results from Proc LOGISTIC. We used the data set talldata provided in the SAS help and documentation to illustrate these comparisons. This dataset contains 34 individuals and 8 markers.

```
proc CASECONTROL data =talldata tall marker=snpname indiv=id;
      var allele1 allele2;
      trait affected;
   run;

We create the variable genotype as follows;

data talldata;
set talldata;
Genotype=compress(allele1|| '/' ||allele2);
If allele1='.' or allele2='.' then genotype='';
run;

proc LOGISTIC data =talldata ;
      class genotype;
      model affected=genotype;
      by Snpname;
      run;

To compare results from trend tests, we create a numerical variable x coded as 0,1,2
(additive allelic model) and we add the information about the major allele and the
minor allele as described in section 1.

data talldata;
set talldata;
if allele1=allele2=majorallele then x=0;
if allele1 ne allele2 then x=1;
if allele1='.' or allele2='.'  then x=.; /*missing value*/
if allele1=allele2=minorallele then x=2;
run;

proc LOGISTIC data =talldata ;
model affected=x;
by Snpname;
run;
```

Table 1 summarizes results obtained with the above three approaches. P-values obtained with both procedures are identical. In particular the Score statistics with proc LOGISTIC is equal to the chi-square statistics of proc CASECONTROL for the genotypic test and the trend test. Based on this comparison, we will use proc LOGISTIC in some situations where proc CASECONTROL can be limited. Examples of these situations are given in the following section.

Table 1: Genotype and trend test results using Proc CASECONTROL and Proc LOGISTIC.

Locus name	Genotypic model					Additive model				
	proc Casecontrol		proc logistic			proc casecontrol		proc logistic		
	ChiSq	p-value	Test	ChiSq	p-value	ChiSq	p-value	Test	ChiSq	p-value
SNP1	0.2724	0.8727	Likelihood Ratio	0.2730	0.8724	0.0319	0.8583	Likelihood Ratio	0.0319	0.8583
SNP1			Score	0.2724	0.8727			Score	0.0319	0.8583
SNP1			Wald	0.2715	0.8731			Wald	0.0319	0.8583
SNP2	3.4298	0.1800	Likelihood Ratio	3.5947	0.1657	2.1397	0.1435	Likelihood Ratio	2.2369	0.1348
SNP2			Score	3.4298	0.1800			Score	2.1397	0.1435
SNP2			Wald	3.1551	0.2065			Wald	2.0218	0.1551

2.3. Situations where Proc CASECONTROL can be limited

2.3.1. Adjusting for a covariate

In genetic association studies, it is often relevant to include known covariates in a model. The CASECONTROL procedure relies primarily on contingency table analysis. To introduce covariates in prediction models one can use regression methods as described in the SAS code below.

```
/*Data one should contain the following variables: id (individual's id), SNPname,
genotype, status (affected/unaffected), covariate (for example, gender or age) and
should be sorted by SNPname */

proc LOGISTIC data=one ;
class genotype  ;
model status=genotype covariate  ;
by SNPname;
run;
```

2.3.2. Testing for gene-gene interaction

In complex traits, multiple genes can interact to produce the disease, and one may detect the joint effect by modeling the complex interaction pattern among loci. The following macro describes the use of Proc LOGISTIC to test for interaction between each pair of SNPs in the dataset:

```
/*Data one should contain the following variables: individualid, SNPname, genotype,
disease_status (affected/unaffected), marker_number (variable that orders SNPs from 1
to the total number of SNPs)*/

proc SORT data=one (rename=(genotype=genotype2 SNPname=snp2)) out=snp2;
by individualid;
run;

%macro interaction;
%let num=&num; /* num is the total number of snps in data one*/
%do i=1 %to &num;
%let marknum=&i;
data snp1 (rename=(genotype=genotype1 SNPname=snp1));
set one(where=( marker_number=%eval(&i)));
run;
data _null_; set snp1 (obs=1); call symput ('first',snp1); run;

/*do a one to many merge*/
data two;
merge snp1 (in=a drop=marker_number) snp2 (in=b where=(marker_number>=%eval(&i)+1));
by individualid;
if a and b;
run;

proc SORT data=two;
by snp2;
run;

ods output parameterestimates=interaction (where =(variable ne 'Intercept'));
ods output convergencestatus=convergence;
proc LOGISTIC data=two;
class genotype1 genotype2 ;
model disease_status =genotype1 genotype2 genotype1*genotype2 ;
by snp2;
run;
ods output close;
ods output close;

data interaction (drop=snp2);
set interaction; snps="&first"||snp2;
marker="&marknum" ;
```

```
run;

data convergence (drop=snp2);
set convergence; snps="&first"||snp2;
marker="&marknum" ;
run;

proc APPEND base=resultint data=interaction;
run;

proc APPEND base= resultconv data=convergence;
run;
%end;
%mend interaction;
%interaction;
```

2.3.3. Testing for gene-environment interaction

Many human diseases appear to be associated with both genetic and environmental factors that are possibly interacting. Once again, Proc LOGISTIC is used instead of Proc CASECONTROL as described below.

```
/*Data one should contain the following variables: id, SNPname, genotype,
environment_status (affected/unaffected) and should be sorted by SNPname*/

proc LOGISTIC data=one ;
class genotype environment ; /* only if the environment is a categorical variable*/
model status=genotype environment genotype*environment ;
by SNPname ;
run;
```

2.3.4. Testing for modes of inheritance

In genetic analysis, it is common to test for model fit. Many different disease models have been suggested for the effect of a single genetic variant, the most common ones are dominant, recessive and additive. The model is said dominant if a single copy of an allele (A for example) is sufficient to cause an increase in disease risk and recessive if two copies are necessary to cause a rise in disease risk. The model is additive or dose-dependent if each allele confers an increase in risk. To test for genotype-disease association according to one of these three models, a single variable, X, can be used in regression equations. Variable X codings for each model are given in Table 1.

Table 2: Numerical coding of genotypes for each mode of inheritance

	X coding of additive model	X coding of dominant model	X coding of recessive model
Common genotype	0	0	0
Heterozygote genotype	1	1	0
Wild genotype	2	1	1

The following code can be used for each of the above described models.

```
/*Data one should contain the following variables: id, status (affected/unaffected),
SNPname, genotype, variable X which represents the numerical coding corresponding to
the mode of inheritance, covariate (if adjusting for any covariate)*/

proc LOGISTIC data=one ;
class covariate ; /* only if the covariate is categorical*/
model status=X covariate;
by SNPname;
run;
```

2.3.5. Testing for population stratification

Population stratification occurs when there are allele frequency differences between cases and controls due to systematic ancestry differences which can cause spurious associations in disease studies. Proc CASECONTROL

provides the genomic control method which corrects for population stratification by adjusting association statistics at each marker by a uniform inflation factor (Devlin & Roeder, 1999). This uniform adjustment may be inappropriate for markers that differ in their allele frequencies across ancestral populations more than others. We have implemented the code of an alternative method called EIGENSTRAT that corrects for population stratification using principal component analysis. For more details see (Price et al, 2005). The SAS code described below uses Proc PRINCOMP, Proc CORR, Proc GPLOT and Proc LOGISTIC.

```
/* Dataset one should contain the following variables: SNPname, id which represents
the individual'id, genotype, variable g (numerical coding of the genotype) which is
coded as 0 if homozygotes for major allele, 1 if heterozygotes, 2 if homozygotes for
minor allele and disease_status (affected /not affected). This dataset should be
sorted by SNPname*/

/* Step1-Write the data as a matrix G=(gij) i=1,...M for SNPs and j=1,...,N for
individuals*/

/*Calculate the mean of each SNP*/
ods output basicmeasures=snpmean(where=(locmeasure='Mean') keep=locmeasure SNPname
locvalue);
proc UNIVARIATE data=one;
var g;
by SNPname;
run;
ods output close;
/* Normalize SNPs*/
data two (keep=SNPname g_normalized id);
merge one snpmean;
by SNPname;
sum=N*locvalue ;/* N is the number of individuals*/
p=(1+sum)/(2+2*N);
variance = p*(1-p);
norme=sqrt(variance);
g_centred=g-locvalue;
g_normalized=g_centred/norme;
run;

/*Construct N*N covariance matrix M of individuals where mjj' is defined to be the
covariance between individual j and individual j'*/
proc TRANSPOSE data=two out=three(drop=SNPname _name_);
var g_normalized;
by SNPname;
id id;
run;

ods listing close;
ods output cov=covariance_matrix (drop=row: col: df:);
proc CORR data=three cov;
run;
ods output close;

/* Step2- Principal Component analysis on Covariance matrix M */
/* The kth axis is the kth eigenvector of M */
ods listing close;
ods output eigenvectors=eigenvectors;
ods output eigenvalues=eigenvalues;
proc PRINCOMP data=covariance_matrix;
run;
ods output close;
ods output close;
ods listing;

/*Merge data one and data eigenvectors*/
proc SORT data=one ;
by id;
run;
```

```
data four;
merge one (keep=id SNPname genotype disease_status) eigenvectors(rename=(variable=id)
keep=prin1 prin2 variable);
by id;
run;

/*Plot of prin1 and prin2 */
symbol1 i=none value=dot color=red;
symbol2 i=none value=dot color=blue;
proc GPLOT data=four;
plot prin2*prin1=disease_status/haxis=axis1 vaxis=axis2;
run;
quit;

/* Correction for stratification using axes of variation (we restrict ourselves here
to the two first axes)*/
proc LOGISTIC data=four ;
class genotype ;
model status=genotype prin1 prin2;
by SNPname;
run;
```

2.3.6. Matched case-control

Matched case-control designs are sometimes used in genetic association studies to conduct case-sibling analysis for example or as a means of adjusting for important covariates such as ethnic origin. Conditional logistic regression is used to investigate the relationship between an outcome and a set of genetic markers (SNPs) in matched case-control studies. The outcome is the disease status. The matching is 1:1 if there is only one case and one control, and the matching is m:n when there is a varying number of cases and controls in the matched sets. Proc PHREG performs conditional logistic regression by forming a stratum for each matched set. The following SAS code shows an example using Proc PHREG with cases and controls matched on age.

```
/*Data one should contain the following variables: id, SNPname, genotype (numerical
variable coded as 0, 1, 2 in this example), status (the variable status is used to
determine the subject is a case (status=1) or control (status=0). The dummy time
variable Time takes the value 1 for cases and 2 for controls. Data one should be
sorted by SNPname */

proc PHREG data=one ;
model time*status(0)=genotype;
strata age;
by SNPname;
run;
```

2.3.7. Survival analysis

Survival analysis involves the modeling of time-to-event data. Subjects are followed until they reach a pre-specified endpoint. Proc PHREG, Proc TPHREG, Proc BPHREG, Proc LIFEREG and Proc LIFETEST are all SAS/STAT procedures that perform analysis of survival data. We provide here an example of using Proc TPHREG in genetic association studies to model the effect of genotypes on hazard rates. (Proc TPHREG allows the addition of the CLASS statement before the MODEL statement.)

```
/* The data set one contains the variable time (the survival time), the variable
censor (the censoring indicator variable 0 if censored and 1 if not censored),
SNPname, individualid and the variable genotype (categorical variable). Data one
should be sorted by SNPname*/
proc TPHREG data=one ;
class genotype ;
model time*censor(0)=genotype;
by SNPname;
run;
```

3. ADJUSTING P-VALUES

Multiple testing is a challenging issue in genetic association studies when a large number of markers are used. Failure to adjust for multiple testing appropriately may produce excessive false positive results. SAS/GENETICS provides the PSMOOTH procedure which relies on smoothing methods that take into account *p*-values from neighboring, and possibly correlated, markers.

3.1. Description of Proc PSMOOTH

The PSMOOTH procedure uses smoothing methods that modify the *p*-value from each marker test using a function of its original *p*-value and the *p*-values of the tests on the nearest markers. Since the number of hypothesis tests being performed is not reduced, adjustments to correct the smoothed *p*-values for multiple testing are required. Bonferroni and Sidak methods are offered by PROC PSMOOTH to adjust the smoothed *p*-values for multiple testing.

3.2. Alternative to Proc PSMOOTH

Recently, Gao et al (2008) proposed a new multiple testing correction *simpleM* which uses composite linkage disequilibrium (CLD) to create the correlation matrix of SNPs and M_{effG} to calculate the effective number of independent tests (for more details, see Gao et al, 2008).

Simple*M* method SAS code

```
/* Step1- Derive the composite LD (CLD) correlation matrix from SNP dataset*/
/* Dataset one contains the SNP genotypes numerically coded as 2,1 and 0 for wild-type
allele homozygotes, heterozygotes and variant-type allele homozygotes, respectively.*/

ods listing close;
ods output pearsoncorr=correlation(drop=variable P: N:);
proc CORR data=one pearson;
run;
ods output close;
ods listing;

/*Step2-Calculate the eigenvalues (principal components analysis)*/
ods output eigenvalues=eigenvalues;
proc PRINCOMP data=correlation ;
run;
ods output close;

/*Step3-Infer MeffG through PCA to estimate the effective number of independent tests*/
Data Meff;
Set eigenvalues;
If cumulative>0.995 then Meff=number; /* cumulative and number are two provided
variables in eigenvalues data set*/
Run;

/*Step4-Apply the Bonferroni correction formula to adjust point-wise significance
level*/

Data Meff;
Set Meff;
alphaG=alphaE/Meff; /*alphaE is the experiment-wise error rate*/
Run;
```

CONCLUSION

SAS GENETICS offers a collection of procedures that are useful for the analysis of genetic data. We have presented an overview of these procedures with examples of complementary SAS/STAT procedures as we routinely use them in the context of human genetic association studies. The SAS code provided within this paper will be useful to researchers actively engaged in genetic studies.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

REFERENCES

- Devlin, B. and K. Roeder, *Genomic control for association studies.* Biometrics, 1999. **55**(4): p. 997-1004.
- Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nat Genet, 2006. **38**(8): p. 904-9.
- Gao, X., J. Starmer, and E.R. Martin, *A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms.* Genet Epidemiol, 2008. **32**(4): p. 361-9.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Marie-Pierre Dubé
Montreal Heart Institute
5000 Belanger
Montreal, QC H1T1C8
Work phone: (514) 376-3330 #2298
Fax: (514) 376-1355
E-mail: marie-pierre.dube@umontreal.ca
Web: www.statgen.org

Amina Barhdadi
Work phone: (514) 376-3330 #3303
E-mail: amina.barhdadi@statgen.org

Yassamin Feroz Zada
Work phone: (514) 376-3330 #3046
E-mail: Yassamin.FerozZada@statgen.org