Chapter 1

What's New in SAS/Genetics 9.0, 9.1, and 9.1.3

Chapter Contents

Chapter 1
What's New in SAS/Genetics 9.0, 9.1, and 9.1.3

Overview

SAS/Genetics includes several new procedures:

- the experimental HTSNP procedure for selecting a subset of SNPs that identify groups of haplotypes that minimize within-group diversity **9.1**
- the INBREED procedure for estimating covariance and/or inbreeding coefficients for pedigrees **9.1**

New features have been added to the SAS/Genetics procedures:

- ALLELE
- CASECONTROL
- FAMILY
- HAPLOTYPE
- PSMOOTH

including options that accommodate new data formats.

Accommodating New Data Formats

There are several new options available for analyzing data in different formats. **9.1** The GENOCOL and DELIMITER= options have been added to four procedures: ALLELE, CASECONTROL, FAMILY, and HAPLOTYPE. The GENOCOL option enables you to use columns containing marker genotypes instead of a pair of columns containing the two alleles that comprise the genotype. You can specify the delimiter that is used to separate the two alleles with the DELIMITER= option. In addition, the experimental options TALL, MARKER=, and INDIV= can be used collectively for data in a "tall-skinny" format in the ALLELE, CASECONTROL, and HAPLOTYPE procedures. Data sets in this format contain a marker identifier and individual identifier, along with one variable containing the marker genotypes or two columns containing marker alleles. See the individual procedures' Syntax sections for more details about these new options.

ALLELE Procedure

9.1 The new options ALLELEMIN=, GENOMIN=, and HAPLOMIN= enable you to specify the minimum estimated frequency for an allele, genotype, or haplotype, respectively, to be included in its corresponding ODS table. By default, any allele, genotype, or haplotype that occurs at least once in the sample is included in the respective table. These options can be used to reduce the size of the ODS tables, or alternatively, the GENOMIN= or HAPLOMIN= options can be set to 0 to view all possible genotypes or haplotypes, not just those that are observed.

Some enhancements have been made in SAS 9.1.3 that improve the performance of calculating linkage disequilibrium (LD) statistics for a large number of marker pairs. These include:

- more efficient usage of memory when calculating LD measures
- a new LOGNOTE option to request notes indicating the status of the LD calculations be printed to the log
- a new WITH statement that gives an alternative way of specifying pairs of markers to analyze for LD and enables the partitioning of these calculations

Beginning in SAS 9.1.3, columns containing counts for alleles, genotypes, and, when HAPLO=GIVEN, haplotypes are included in the corresponding ODS table in addition to the relative frequencies that have always been displayed. Cosmetic improvements to the ODS tables in the SAS listing make these tables easier to read.

Also new, the HAPLO=NONEHWD option can be used to request that the CLD test statistic for biallelic markers be adjusted for Hardy-Weinberg disequilibrium (Weir 1996). This adjustment is also made in the correlation coefficient when requested to be displayed in the "LD Measures" table.

CASECONTROL Procedure

9.1 The new NULLSNPS= option enables you to specify SNPs to be used in calculating the variance inflation factor for genomic control. By default, if VIF is specified, the variables in the VAR statement are used, but this new option provides a way of using particular SNPs, separate from those being tested for association and which are assumed to have no association with the TRAIT variable, for genomic control (Bacanu, Devlin, and Roeder 2000).

9.1 You can request that approximations of exact p-values for the case-control tests be reported in place of the asymptotic chi-square p-values (Westfall and Young 1993). The new PERMS= option indicates the number of permutations to be used for a Monte Carlo estimate of each exact p-value, and the random seed can be provided in the new SEED= option.

9.1 The OUTSTAT= data set includes two new columns: NumTrait1 and NumTrait2, where the values 1 and 2 are replaced by the two values of the TRAIT variable.

These columns contain the number of genotyped individuals with each trait value for each marker.

Beginning in SAS 9.1.3, the STRATA statement is available for specifying variables that define strata in the sample. A Cochran-Mantel-Haenszel statistic (Agresti 1990) is used to adjust for these categorical variables, enabling you to stratify your analysis on the basis of gender or treatment for example, or to perform a nested or matched case-control study.

Also new to SAS 9.1.3 is an OR option to have allele odds ratios and their confidence intervals, with the level specified in the ALPHA= option, included in the output data set for biallelic markers.

FAMILY Procedure

The new "Family Summary" ODS table displays information about each family in the data set at each of the markers. This includes the number of parents genotyped, the number of affected children, the number of unaffected children, as well as an error code indicating what type of, if any, Mendelian inconsistencies occur in a nuclear family's genotypes at each marker. The new SHOWALL option can be used to display this information for all families at each marker. By default, only those families with a genotype error are included in the table for the marker(s) where the error occurs. ***9.1***

The new "Description of Error Codes" ODS table provides descriptions of the numerical error codes used in the "Family Summary" table. ***9.1***

Approximations of exact p-values can now be requested in place of the asymptotic chi-square p-values for the TDT, S-TDT, SDT, and combined S-TDT and SDT using the PERMS= option. The number specified indicates the number of permutations to be used in the Monte Carlo procedure for estimating exact p-values. You can provide the random seed used for the permutations in the new SEED= option. ***9.1***

The multiallelic SDT and multiallelic combined SDT/TDT are now implemented as described by Czika and Berry (2002). ***9.1***

Analysis of X-linked markers is facilitated in SAS 9.1.3 by the new XLVAR statement. Markers listed in this statement can be tested for linkage and, under appropriate conditions, association with a binary trait using the X-linked versions of the TDT, S-TDT, combined S-TDT, and RC-TDT (Horvath, Laird, and Knapp 2000).

With the OUTQ= option, new to SAS 9.1.3 as well, you can create an output data set containing the pair of allelic transmission scores (Abecasis, Cookson, and Cardon 2000) for each marker allele. These scores can then be used to perform family-based tests for binary or quantitative traits in nuclear families or even extended pedigrees.

HAPLOTYPE Procedure

9.1 The EST=EM | STEPEM option enables you to specify whether you would like haplotype frequencies to be estimated using the original EM algorithm or the new stepwise EM algorithm (Clayton 2002b). When EST=STEPEM is specified, a cutoff to be used for trimming the set of haplotypes before adding an additional locus can be given in the new STEPTRIM= option.

9.1 The ID statement enables variables from the input data set to be included in the OUT= data set created by PROC HAPLOTYPE in addition to or instead of the _ID_ variable, a unique numeric identifier assigned to each individual by the procedure.

The option EST=BAYESIAN is experimental in SAS 9.1.3 and can be used to request a Bayesian approach to the estimation of haplotype frequencies. Some new options associated with only this estimation method are BURNIN=, INTERVAL=, THETA=, and TOTALRUN=, all experimental as well.

HTSNP Procedure

9.1 The experimental HTSNP procedure implements search algorithms for identifying a subset of SNPs called *haplotype tag SNPs (htSNPs)* (Johnson et al. 2001) that capture much of the linkage disequilibrium and haplotype diversity among common haplotypes.

Beginning in SAS 9.1.3, PROC HTSNP contains an option CRITERION= for selecting the measure to be used for determining the best set(s) of haplotype-tagging SNPs (htSNPs). The possible values for this option are PDE (Clayton 2002a), the default and the measure previously available, and RSQH, which implements the R_h^2 measure of Stram et al. (2003). Additionally, the best sets of htSNPs and their criterion measure are now displayed automatically in the ODS table "Evaluation of htSNPs" so the OUTSTAT= option is no longer offered.

INBREED Procedure

9.1 The INBREED procedure is now included in SAS/Genetics in addition to SAS/STAT where it originated. This procedure calculates the covariance or inbreeding coefficients for pedigrees either by treating the population as a single generation or by performing separate analyses on each generation. You can also opt to have inbreeding and covariance coefficients averaged within each gender category.

PSMOOTH Procedure

9.1

The new option TPM implements the truncated product method (Zaykin et al. 2002) for smoothing p-values over windows of markers. The TAU= option, also new, can be used in conjunction with the TPM option to specify the value of τ at which p-values are truncated.

The Benjamini and Hochberg (1995) method of adjusting p-values to control the false discovery rate (FDR) is available in SAS 9.1.3 with the ADJUST=FDR option.

References

Abecasis, G.R., Cookson, W.O.C., and Cardon, L.R. (2000), "Pedigree Tests of Transmission Disequilibrium," *European Journal of Human Genetics,* 8, 545–551.

Agresti, A. (1990), *Categorical Data Analysis,* New York: John Wiley & Sons, Inc.

Bacanu, S-A., Devlin, B., and Roeder, K. (2000), "The Power of Genomic Control," *American Journal of Human Genetics,* 66, 1933–1944.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society, Series B,* 57, 289–300.

Clayton, D. (2002a), "Choosing a Set of Haplotype Tagging SNPs from a Larger Set of Diallelic Loci," [http://www-gene.cimr.cam.ac.uk/clayton/software/stata/htSNP/htsnp.pdf].

Clayton, D. (2002b), "SNPHAP: A Program for Estimating Frequencies of Large Haplotypes of SNPs," [http://www-gene.cimr.cam.ac.uk/clayton/software/snphap.txt].

Czika, W. and Berry, J.J. (2002), "Using All Alleles in the Multiallelic Versions of the SDT and Combined SDT/TDT," *American Journal of Human Genetics,* 71, 1235–1236.

Horvath, S., Laird, N.M., and Knapp, M. (2000), "The Transmission/Disequilibrium Test and Parental-Genotype Reconstruction for X-Chromosomal Markers," *American Journal of Human Genetics,* 66, 1161–1167.

Johnson, G.C.L. et al. (2001), "Haplotype Tagging for the Identification of Common Disease Genes," *Nature Genetics,* 29, 233–237.

Stram, D.O., Haiman, C.A., Hirschhorn, D.A., Kolonel, L.N., Henderson, B.E., and Pike, M.C. (2003), "Choosing Haplotype-Tagging SNPs Based on Unphased Genotype Data Using a Preliminary Sample of Unrelated Subjects with an Example from the Multiethnic Cohort Study," *Human Heredity,* 55, 27–36.

Weir, B.S. (1996), *Genetic Data Analysis II,* Sunderland, MA: Sinauer Associates, Inc.

Westfall, P.H. and Young, S.S. (1993), *Resampling-based Multiple Testing,* New York: John Wiley & Sons, Inc.

Zaykin, D.V., Zhivotovsky, L.A., Westfall, P.H., and Weir, B.S. (2002), "Truncated Product Method for Combining P-values," *Genetic Epidemiology,* 22, 170–185.