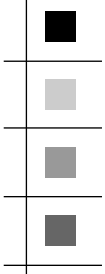
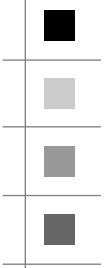
SAS/Genetics® User's Guide



The Power to Know™

SAS/Genetics® User's Guide

References to this documentation being online with the software should be ignored. The documentation is actually online at www.sas.com/service/library/onlinedoc/genetics.



The Power to Know™

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS/Genetics® User's Guide,* Cary, NC: SAS Institute Inc., 2002.

SAS/Genetics® User's Guide

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

51223_0502

Contents

Credits

Documentation

Writing	Wendy Czika and Xiang Yu
Editing	Virginia Clark and Rob Pratt
Technical Review	Russell D. Wolfinger and Robert N. Rodriguez
Documentation Production	Tim Arnold

Software

Procedures	Wendy Czika and Xiang Yu
TPLOT Results	Art Barnes and Susan R. Edwards

Support Groups

Quality Assurance	Ming-Chun Chang, Bruce Elsheimer, Gregory D. Goodwin, Kelly M. Graham, Gerardo I. Hurtado, Bob Overby, Greg Weier, and Charlotte Zinkus
Technical Support	Rob Agnelli and Kathleen Kiernan

Acknowledgments

Many people make significant and continuing contributions to the development of SAS software products. The following are some of those who have contributed significant amounts of their time to help us make improvements to SAS/Genetics software. This includes research and consulting, testing, or reviewing documentation. We are grateful for the involvement of these members of the statistical community and the many others who are not mentioned here for their feedback, suggestions, and consulting.

Margaret G. Ehm	GlaxoSmithKline
Greg Gibson	North Carolina State University
Shizue Izumi	Radiation Effects Research Foundation
Dahlia Nielsen	North Carolina State University
Bruce S. Weir	North Carolina State University
Dmitri Zaykin	GlaxoSmithKline

The simulated GAW12 data used in this book are supported by NIGMS grant GM31575.

The final responsibility for the SAS System lies with SAS alone. We hope that you will always let us know your opinions about the SAS System and its documentation. It is through your participation that SAS Software is continuously improved.

Please send your comments to **suggest@sas.com**.

Chapter 1
Introduction

Chapter Contents

Chapter 1
Introduction

Overview of SAS/Genetics Software

Statistical analyses of genetic data are now central to medicine, agriculture, evolutionary biology, and forensic science. The inherent variation in genetic data, together with the substantial increase in the scale of genetic data following the human genome project, has created a need for reliable computer software to perform these analyses. The procedures offered by SAS/Genetics and described here represent an initial response of SAS Institute to this need.

Although many of the statistical techniques used in the new procedures are standard, others have had to be developed to reflect the genetic nature of the data. All the procedures are designed to operate on data sets that have a familiar structure to geneticists, and that mirror those used in existing software. The syntax for these genetic analyses follows that familiar to SAS users, and the output can be tabular or graphical. The objective of the procedures is to bring the full power of SAS analyses to bear on the characterization of fundamental genetic parameters, and most importantly on the detection of associations between genetic markers and disease status.

Most of the analyses in SAS/Genetics are concerned with detecting patterns of covariation in genetic marker data. These data generally consist of pairs of discrete categories; this pairing derives from the underlying biology, namely the fact that complex organisms have pairs of chromosomes. Each marker refers to the genetic status of a *locus,* each marker type is called an *allele,* and each pair of alleles in an individual is called a *genotype.* A set of alleles present on a single chromosome is called a *haplotype.* Genetic markers may be single nucleotide polymorphisms (SNPs), which are sites in the DNA where the nucleotide varies among individuals, usually with only two alleles possible; microsatellites, which are simple sequence repeats that generate usually between 2 and 20 categories; and other classes of DNA variation.

Two of the procedures in SAS/Genetics are concerned solely with the analysis of genetic marker data. The ALLELE procedure calculates descriptive statistics such as the frequency and variance of alleles and genotypes, as well as estimating measures of marker informativeness, and testing whether genotype frequencies are consistent with Hardy-Weinberg equilibrium (HWE). This procedure also supports three methods for calculation of the degree and significance of *linkage disequilibrium* (LD) among markers at pairs of loci, where LD refers to the propensity of alleles to cosegregate. The HAPLOTYPE procedure is used to infer the most likely multilocus haplotype frequencies in a set of genotypes. Since genetic markers are usually measured independently of one another, there is no direct way to determine which two alleles were on the same chromosome. The algorithm implemented in this procedure converges on the haplotype frequencies that have the highest probability of generating the observed genotypes.

Many genetic data sets are now used to study the relationship between genetic markers and complex phenotypes, particularly disease susceptibility. In general terms, traits can be measured as continuous variables (for example, weight or serum glucose concentration), as discrete numerical categories (for example, meristic measures or psychological class), or as affected/unaffected indicator variables. The two procedures CASECONTROL and FAMILY both take simple dichotomous indicators of disease status and use standard genetic algorithms to compute statistics of association between these indicators and the genetic markers. The CASECONTROL procedure is designed to contrast allele and genotype frequencies between affected and unaffected populations, using three types of chi-square tests and options for controlling correlation of allele frequencies among members of the same subpopulation. Significant associations may indicate that the marker is linked to a locus that contributes to disease susceptibility, though population structure in conjunction with environmental or cultural variables can also lead to associations, and the statistical results must be interpreted with caution. The FAMILY procedure employs several transmission/disequilibrium tests of nonrandom association between disease status and linkage to markers transmitted from heterozygous parents to affected offspring (TDT) or pairs of affected and unaffected siblings (S-TDT and SDT). A joint analysis known as the reconstruction-combined TDT (RC-TDT) can also accommodate missing parental genotypes and families lacking unaffected children under some circumstances.

The output of these procedures can be further explored by using the PSMOOTH procedure to adjust p-values from association tests performed on large numbers of markers obtained in a genome scan, or by creating a graphical representation of the procedures' output, namely p-values from tests for LD, HWE, and marker-disease associations, using the %TPLOT macro.

About This Book

Since SAS/Genetics software is a part of the SAS System, this book assumes that you are familiar with base SAS software and with the books *SAS Language Reference: Dictionary*, *SAS Language Reference: Concepts*, and the *SAS Procedures Guide*. It also assumes that you are familiar with basic SAS System concepts such as creating SAS data sets with the DATA step and manipulating SAS data sets with the procedures in base SAS software (for example, the PRINT and SORT procedures).

Chapter Organization

This book is organized as follows.

Chapter 1, this chapter, provides an overview of SAS/Genetics software and summarizes related information, products, and services. The next five chapters describe the SAS procedures that make up SAS/Genetics software. These chapters appear in alphabetical order by procedure name. They are followed by a chapter documenting a SAS macro provided with SAS/Genetics software.

The chapters documenting the SAS/Genetics procedures are organized as follows:

- The *Overview* section provides a brief description of the analysis provided by the procedure.
- The *Getting Started* section provides a quick introduction to the procedure through a simple example.
- The *Syntax* section describes the SAS statements and options that control the procedure.
- The *Details* section discusses methodology and miscellaneous details.
- The *Examples* section contains examples using the procedure.
- The *References* section contains references for the methodology and examples for the procedure.

Typographical Conventions

This book uses several type styles for presenting information. The following list explains the meaning of the typographical conventions used in this book:

roman	is the standard type style used for most text.
UPPERCASE ROMAN	is used for SAS statements, options, and other SAS language elements when they appear in the text. However, you can enter these elements in your own SAS programs in lowercase, uppercase, or a mixture of the two.
UPPERCASE BOLD	is used in the "Syntax" sections' initial lists of SAS statements and options.
oblique	is used for user-supplied values for options in the syntax definitions. In the text, these values are written in *italic*.
helvetica	is used for the names of variables and data sets when they appear in the text.
bold	is used to refer to matrices and vectors.
italic	is used for terms that are defined in the text, for emphasis, and for references to publications.
`monospace`	is used for example code. In most cases, this book uses lowercase type for SAS code.

Options Used in Examples

Output of Examples

For each example, the procedure output is numbered consecutively starting with 1, and each output is given a title. Each page of output produced by a procedure is enclosed in a box. Most of the output shown in this book is produced with the following SAS System options:

```
options linesize=80 pagesize=200 nonumber nodate;
```

In some cases, if you run the examples, you will get slightly different output depending on the SAS system options you use and the precision used for floating-point calculations by your computer. This does not indicate a problem with the software. In all situations, any differences should be very small.

Graphics Options

The examples that contain graphical output are created with a specific set of options and symbol statements. The code you see in the examples creates the color graphics that appear in the online (CD) version of this book. A slightly different set of options and statements is used to create the black-and-white graphics that appear in the printed version of the book.

If you run the examples, you may get slightly different results. This may occur because not all graphic options for color devices translate directly to black-and-white output formats. For complete information on SAS/GRAPH software and graphics options, refer to *SAS/GRAPH Software: Reference*.

The following GOPTIONS statement is used to create the online (color) version of the graphic output.

```
filename GSASFILE  '<file-specification>';

goptions gsfname=GSASFILE   gsfmode =replace
         fileonly
         transparency       dev     = gif
         ftext   = swiss    lfactor = 1
         htext   = 4.0pct   htitle  = 4.5pct
         hsize   = 5.625in  vsize   = 3.5in
         noborder           cback   = white
         horigin = 0in      vorigin = 0in ;
```

The following GOPTIONS statement is used to create the black-and-white version of the graphic output, which appears in the printed version of the manual.

```
filename GSASFILE  '<file-specification>';

goptions gsfname=GSASFILE   gsfmode =replace
         gaccess = sasgaedt fileonly
         dev     = pslepsf
         ftext   = swiss    lfactor = 1
         htext   = 3.0pct   htitle  = 3.5pct
         hsize   = 5.625in  vsize   = 3.5in
         border             cback   = white
         horigin = 0in      vorigin = 0in ;
```

In most of the online examples, the plot symbols are specified as follows:

```
symbol1 value=dot color=white height=3.5pct;
```

The SYMBOL*n* statements used in online examples order the symbol colors as follows: white, yellow, cyan, green, orange, blue, and black.

In the examples appearing in the printed manual, symbol statements specify COLOR=BLACK and order the plot symbols as follows: dot, square, triangle, circle, plus, x, diamond, and star.

Where to Turn for More Information

This section describes other sources of information about SAS/Genetics software.

Online Help System

You can access online help information about SAS/Genetics software in two ways. You can select **SAS System Help** from the **Help** pull-down menu and then select **SAS/Genetics Software** from the list of available topics. Or, you can bring up a command line and issue the command **help Genetics** to bring up an index to the statistical procedures, or issue the command **help ALLELE** (or another procedure name) to bring up the help for that particular procedure. Note that the online help includes syntax and some essential overview and detail material.

SAS Institute Technical Support Services

As with all SAS Institute products, the SAS Institute Technical Support staff is available to respond to problems and answer technical questions regarding the use of SAS/Genetics software.

Related SAS Software

Many features not found in SAS/Genetics software are available in other parts of the SAS System. If you do not find something you need in SAS/Genetics software, try looking for the feature in the following SAS software products.

Base SAS Software

The features provided by SAS/Genetics software are in addition to the features provided by Base SAS software. Many data management and reporting capabilities you will need are part of Base SAS software. Refer to *SAS Language Reference: Concepts*, *SAS Language Reference: Dictionary*, and the *SAS Procedures Guide* for documentation of Base SAS software.

SAS DATA Step

The DATA step is your primary tool for reading and processing data in the SAS System. The DATA step provides a powerful general-purpose programming language that enables you to perform all kinds of data processing tasks. The DATA step is documented in *SAS Language Reference: Concepts*.

Base SAS Procedures

Base SAS software includes many useful SAS procedures. Base SAS procedures are documented in the *SAS Procedures Guide*. The following is a list of Base SAS procedures you may find useful:

CHART	for printing charts and histograms
CONTENTS	for displaying the contents of SAS data sets
CORR	for computing correlations
FREQ	for computing frequency crosstabulations
MEANS	for computing descriptive statistics and summarizing or collapsing data over cross sections
PRINT	for printing SAS data sets
SORT	for sorting SAS data sets
TABULATE	for printing descriptive statistics in tabular format
TRANSPOSE	for transposing SAS data sets
UNIVARIATE	for computing descriptive statistics

SAS/GRAPH Software

SAS/GRAPH software includes procedures that create two- and three-dimensional high-resolution color graphics plots and charts. You can generate output that graphs the relationship of data values to one another, enhance existing graphs, or simply create graphics output that is not tied to data.

SAS/IML Software

SAS/IML software gives you access to a powerful and flexible programming language (Interactive Matrix Language) in a dynamic, interactive environment. The fundamental object of the language is a data matrix. You can use SAS/IML software interactively (at the statement level) to see results immediately, or you can store statements in a module and execute them later. The programming is dynamic because necessary activities such as memory allocation and dimensioning of matrices are done automatically. SAS/IML software is of interest to users of SAS/Genetics software because it enables you to program your own methods in the SAS System.

SAS/INSIGHT Software

SAS/INSIGHT software is a highly interactive tool for data analysis. You can explore data through a variety of interactive graphs including bar charts, scatter plots, box plots, and three-dimensional rotating plots. You can examine distributions and perform parametric and nonparametric regression, analyze general linear models and generalized linear models, examine correlation matrices, and perform principal component analyses. Any changes you make to your data show immediately in all graphs

and analyses. You can also configure SAS/INSIGHT software to produce graphs and analyses tailored to the way you work.

SAS/INSIGHT software may be of interest to users of SAS/Genetics software for interactive graphical viewing of data, editing data, exploratory data analysis, and checking distributional assumptions.

SAS/STAT Software

SAS/STAT software includes procedures for a wide range of statistical methodologies including

- logistic and linear regression
- censored regression
- principal component analysis
- variance component analysis
- cluster analysis
- contingency table analysis
- categorical data analysis: log-linear and conditional logistic models
- general linear models
- linear and nonlinear mixed models
- generalized linear models
- multiple hypothesis testing

SAS/STAT software is of interest to users of SAS/Genetics software because many statistical methods for analyzing genetics data not included in SAS/Genetics software are provided in SAS/STAT software.

Chapter 2
The ALLELE Procedure

Chapter Contents

Chapter 2
The ALLELE Procedure

Overview

The ALLELE procedure performs preliminary analyses on genetic marker data. These analyses serve to characterize the markers themselves or the population from which they were sampled, and can also serve as the basis for joint analyses on markers and traits. A *genetic marker* is any heritable unit that obeys the laws of transmission genetics, and the analyses presented here assume the marker genotypes are determined without error. With an underlying assumption of random sampling, the analyses rest on the multinomial distribution of marker alleles, and many standard statistical techniques can be invoked with little modification. The ALLELE procedure uses the notation and concepts described by Weir (1996); this is the reference for all equations and methods not otherwise cited.

Data are usually collected at the genotypic level but interest is likely to be centered on the constituent alleles, so the first step is to construct tables of allele and genotype frequencies. When alleles are independent within individuals, that is when there is Hardy-Weinberg equilibrium (HWE), analyses can be conducted at the allelic level. For this reason the ALLELE procedure allows for Hardy-Weinberg testing, although testing is also recommended as a means for detecting possible errors in data.

PROC ALLELE calculates the PIC, heterozygosity, and allelic diversity measures that serve to give an indication of marker informativeness. Such measures can be useful in determining which markers to use for further linkage or association testing with a trait. High values of these measures are a sign of marker informativeness, which is a desirable property in linkage and association tests.

Associations between markers may also be of interest. PROC ALLELE provides tests and various statistics for the association, also called the linkage disequilibrium, between each pair of markers. These statistics can be formed either by using haplotypes that are given in the data, by estimating the haplotype frequencies, or by using only genotypic information.

Getting Started

Example

Suppose you have genotyped 25 individuals at five markers. You want to examine some basic properties of these markers, such as whether they are in HWE, how many alleles each has, what genotypes appear in the data, and if there is linkage disequilibrium between any pairs of markers. You have ten columns of data, with the first two columns containing the set of alleles at the first marker, the next two columns containing the set of alleles for the second marker, and so on. There is one row per each individual. You input your data as follows:

```
data markers;
   input (m1-m10) ($);
   datalines;
B  B  A  B  B  B  A  A  B  B
A  A  B  B  A  B  A  B  C  C
B  B  A  A  B  B  B  B  A  C
A  B  A  B  A  B  A  B  A  B
A  A  A  B  A  B  B  B  C  C
B  B  A  A  A  B  A  B  C  C
A  B  B  B  A  B  A  A  A  B
A  B  A  A  A  A  A  A  A  A
B  B  A  A  A  A  A  B  B  B
A  B  A  B  A  B  B  B  A  C
A  A  A  B  A  A  A  B  B  C
B  B  A  B  A  B  A  B  A  C
A  B  B  B  A  A  A  B  A  C
B  B  B  B  A  A  A  A  A  B
A  B  A  A  A  B  A  A  C  C
A  B  A  A  A  B  A  B  C  C
B  B  A  A  A  A  A  B  A  A
A  A  A  B  A  A  A  B  A  B
A  B  A  A  A  A  B  B  C  C
A  A  A  A  A  A  A  A  B  B
A  B  B  B  A  A  A  A  C  C
A  B  A  B  A  B  A  A  B  B
B  B  A  B  A  B  A  A  A  C
A  B  A  A  A  B  A  B  A  C
A  B  B  B  B  B  A  B  B  B
;
```

You can now use PROC ALLELE to examine the frequencies of alleles and genotypes in your data, and see if these frequencies are occurring in proportions you would expect. The following statements will perform the analysis you want:

```
proc allele data=markers outstat=ld  prefix=Marker
            exact=10000 boot=1000 seed=123;
   var m1-m10;
run;

proc print data=ld;
run;
```

This analysis is using 10,000 permutations to approximate an exact p-value for the HWE test as well as 1,000 bootstrap samples to obtain the confidence interval for the allele frequencies and one-locus Hardy-Weinberg disequilibrium (HWD) coefficients. The starting seed for the random number generator is 123. The PREFIX= option requests that the five markers be named Marker1–Marker5. Since the BOOTSTRAP= option is specified but the ALPHA= option is omitted, a 95% confidence interval is calculated by default.

All five markers are included in the analysis since the ten variables containing the alleles for those five markers were specified in the VAR statement.

The results from the analysis are as follows.

The ALLELE Procedure

Marker Summary

Locus	Number of Indiv	Number of Alleles	PIC	Hetero-zygosity	Allelic Diversity	Test for HWE: Chi-Square	DF	Pr > ChiSq	Prob Exact
Marker1	25	2	0.3714	0.4800	0.4928	0.0169	1	0.8967	1.0000
Marker2	25	2	0.3685	0.3600	0.4872	1.7041	1	0.1918	0.2262
Marker3	25	2	0.3546	0.4800	0.4608	0.0434	1	0.8350	1.0000
Marker4	25	2	0.3648	0.4800	0.4800	0.0000	1	1.0000	1.0000
Marker5	25	3	0.5817	0.4400	0.6552	9.3537	3	0.0249	0.0106

Figure 2.1. Marker Summary for the ALLELE Procedure

Figure 2.1 displays information about the five markers. From this output, you can conclude that Marker5 is the only one showing significant departure from HWE.

Allele Frequencies

Locus	Allele	Frequency	Standard Error	95% Confidence Limits	
Marker1	A	0.4400	0.0711	0.3000	0.5800
Marker1	B	0.5600	0.0711	0.4200	0.7000
Marker2	A	0.5800	0.0784	0.4200	0.7400
Marker2	B	0.4200	0.0784	0.2600	0.5800
Marker3	A	0.6400	0.0665	0.5200	0.7600
Marker3	B	0.3600	0.0665	0.2400	0.4800
Marker4	A	0.6000	0.0693	0.4600	0.7400
Marker4	B	0.4000	0.0693	0.2600	0.5400
Marker5	A	0.2800	0.0637	0.1400	0.4200
Marker5	B	0.3000	0.0800	0.1600	0.4600
Marker5	C	0.4200	0.0833	0.2800	0.6000

Figure 2.2. Allele Frequencies for the ALLELE Procedure

Figure 2.2 displays the allele frequencies for each marker with their standard errors and the lower and upper limits of the 95% confidence interval.

Genotype Frequencies

Locus	Genotype	Frequency	HWD Coeff	Standard Error	95% Confidence Limits	
Marker1	A/A	0.2000	0.006	0.0493	-0.092	0.096
Marker1	A/B	0.4800	0.006	0.0493	-0.092	0.096
Marker1	B/B	0.3200	0.006	0.0493	-0.092	0.096
Marker2	A/A	0.4000	0.064	0.0477	-0.034	0.148
Marker2	A/B	0.3600	0.064	0.0477	-0.034	0.148
Marker2	B/B	0.2400	0.064	0.0477	-0.034	0.148
Marker3	A/A	0.4000	-0.010	0.0457	-0.104	0.080
Marker3	A/B	0.4800	-0.010	0.0457	-0.104	0.080
Marker3	B/B	0.1200	-0.010	0.0457	-0.104	0.080
Marker4	A/A	0.3600	0.000	0.0480	-0.092	0.086
Marker4	A/B	0.4800	0.000	0.0480	-0.092	0.086
Marker4	B/B	0.1600	0.000	0.0480	-0.092	0.086
Marker5	A/A	0.0800	0.002	0.0405	-0.076	0.082
Marker5	A/B	0.1600	0.004	0.0337	-0.066	0.064
Marker5	A/C	0.2400	-0.002	0.0380	-0.074	0.068
Marker5	B/B	0.2000	0.110	0.0445	0.014	0.188
Marker5	B/C	0.0400	0.106	0.0282	0.044	0.156
Marker5	C/C	0.2800	0.104	0.0453	0.010	0.188

Figure 2.3. Genotype Frequencies for the ALLELE Procedure

Figure 2.3 displays the genotype frequencies for each marker with the associated disequilibrium coefficient, its standard error, and the 95% confidence limits.

Obs	Locus1	Locus2	NIndiv	Test	ChiSq	DF	ProbChi	ProbEx
1	Marker1	Marker1	25	HWE	0.01687	1	0.89667	1.0000
2	Marker1	Marker2	25	LD	1.05799	1	0.30367	0.6707
3	Marker1	Marker3	25	LD	1.42074	1	0.23328	0.6524
4	Marker1	Marker4	25	LD	0.33144	1	0.56481	0.9668
5	Marker1	Marker5	25	LD	2.29785	2	0.31698	0.8398
6	Marker2	Marker2	25	HWE	1.70412	1	0.19175	0.2262
7	Marker2	Marker3	25	LD	0.13798	1	0.71030	0.7242
8	Marker2	Marker4	25	LD	1.34100	1	0.24686	0.9015
9	Marker2	Marker5	25	LD	1.13574	2	0.56673	0.5503
10	Marker3	Marker3	25	HWE	0.04340	1	0.83497	1.0000
11	Marker3	Marker4	25	LD	0.46296	1	0.49624	0.9323
12	Marker3	Marker5	25	LD	0.95899	2	0.61909	0.2624
13	Marker4	Marker4	25	HWE	0.00000	1	1.00000	1.0000
14	Marker4	Marker5	25	LD	6.16071	2	0.04594	0.9235
15	Marker5	Marker5	25	HWE	9.35374	3	0.02494	0.0106

Figure 2.4. Testing for Disequilibrium Using the ALLELE Procedure

Figure 2.4 displays the output data set created using the OUTSTAT= option of the PROC ALLELE statement. This data set contains the statistics for testing individual markers for HWE and marker pairs for linkage disequilibrium.

Syntax

The following statements are available in PROC ALLELE.

PROC ALLELE < *options* > **;**
BY *variables* **;**
VAR *variables* **;**

Items within angle brackets (< >) are optional, and statements following the PROC ALLELE statement can appear in any order. The VAR statement is required. The syntax of each statement is described in the following section in alphabetical order after the description of the PROC ALLELE statement.

PROC ALLELE Statement

PROC ALLELE < *options* > **;**

You can specify the following options in the PROC ALLELE statement.

ALPHA=*number*
specifies that a confidence level of $100(1-$*number*$)\%$ is to be used in forming bootstrap confidence intervals for estimates of allele frequencies and disequilibrium coefficients. The value of *number* must be between 0 and 1, and is set to 0.05 by default.

BOOTSTRAP=*number*
BOOT=*number*
indicates that bootstrap confidence intervals should be formed for the estimates of allele frequencies and one-locus disequilibrium coefficients using *number* random samples. One thousand samples are usually recommended to form confidence intervals. If this statement is omitted, no confidence limits are reported.

CORRCOEFF
requests that the "Linkage Disequilibrium Measures" table be displayed and contain the correlation coefficient r, a linkage disequilibrium measure.

DATA=*SAS-data-set*
names the input SAS data set to be used by PROC ALLELE. The default is to use the most recently created data set.

DELTA
requests that the "Linkage Disequilibrium Measures" table be displayed and contain the population attributable risk δ, a linkage disequilibrium measure. This option is ignored if HAPLO=NONE.

DPRIME
requests that the "Linkage Disequilibrium Measures" table be displayed and contain Lewontin's D', a linkage disequilibrium measure.

EXACT=*number*

indicates that the exact p-values for the disequilibrium tests should be calculated using *number* permutations. Large values of *number* (10,000 or more) are usually recommended for accuracy, but long execution times may result, particularly with large data sets. When this option is omitted, no exact tests are performed. If HAPLO=EST, then only the exact tests for Hardy-Weinberg equilibrium are performed; the exact tests for linkage disequilibrium cannot be performed since haplotypes are unknown.

HAPLO=*NONE* **|** *EST* **|** *GIVEN*

indicates whether haplotypes frequencies should not be used, haplotype frequencies should be estimated, or observed haplotype frequencies in the data should be used. This option affects all linkage disequilibrium tests and measures. By default or when HAPLO=NONE is specified, the composite linkage disequilibrium (CLD) coefficient is used in place of the usual linkage disequilibrium (LD) coefficient. In addition, the composite haplotype frequencies are used to form the linkage disequilibrium measures indicated by the options CORRCOEFF and DPRIME. When HAPLO=EST, the maximum likelihood estimates of the haplotype frequencies are used to calculate the LD test statistic as well as the LD measures. The HAPLO=GIVEN option indicates that the haplotypes have been observed, and thus the observed haplotype frequencies are used in the LD test statistic and measures.

When HAPLO=GIVEN, haplotypes are denoted in the data with all alleles comprising one of an individual's two haplotypes in the first of the two variables listed for each marker, and alleles of the other haplotype in the second of the two variables listed for each marker.

MAXDIST=*number*

specifies the maximum number of markers apart that a pair of markers can be in order to perform any linkage disequilibrium calculations. For example, if MAXDIST=1 is specified, linkage disequilibrium measures and statistics are calculated only for pairs of markers that are one apart, such as M1 and M2, M2 and M3, and so on. The number specified must be an integer and is set to 50 markers by default. This option assumes that markers are specified in the VAR statement in the physical order in which they appear on a chromosome or across the genome.

NDATA=*SAS-data-set*

names the input SAS data set containing names, or identifiers, for the markers used in the output. There must be a NAME variable in this data set, which should contain the same number of rows as there are markers in the input data set specified in the DATA= option. When there are fewer rows than there are markers, markers without a name are named using the PREFIX= option. Likewise, if there is no NDATA= data set specified, the PREFIX= option is used.

NOFREQ

suppresses the display of the "Allele Frequencies" and the "Genotype Frequencies" tables. See the section "Displayed Output" on page 26 for a detailed description of these tables.

NOPRINT
suppresses the normal display of results. Note that this option temporarily disables the Output Delivery System (ODS).

OUTSTAT=*SAS-data-set*
names the output SAS data set containing the disequilibrium statistics, for both within-marker and between-marker disequilibria.

PREFIX=*prefix*
specifies a prefix to use in constructing names for marker variables in all output. For example, if PREFIX=VAR, the names of the variables are VAR1, VAR2, ..., VAR*n*. Note that this option is ignored when the NDATA= option is specified, unless there are fewer names in the NDATA data set than there are markers. If this option is omitted, PREFIX=M is the default.

PROPDIFF
requests that the "Linkage Disequilibrium Measures" table be displayed and contain the proportional difference d, a linkage disequilibrium measure. This option is ignored if HAPLO=NONE.

SEED=*number*
specifies the initial seed for the random number generator used for permuting the data in the exact tests and for the bootstrap samples. The value for *number* must be a positive integer; the computer clock time is the default. For more details about seed values, refer to *SAS Language Reference: Concepts*.

YULESQ
requests that the "Linkage Disequilibrium Measures" table be displayed and contain Yule's Q, a linkage disequilibrium measure. This option is ignored if HAPLO=NONE.

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC ALLELE to obtain separate analyses on observations in groups defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in the order of the BY variables. The *variables* are one or more variables in the input data set.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data using the SORT procedure with a similar BY statement.
- Specify the BY statement option NOTSORTED or DESCENDING in the BY statement for the ALLELE procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables using the DATASETS procedure (in Base SAS software).

For more information on the BY statement, refer to the discussion in *SAS Language Reference: Concepts*. For more information on the DATASETS procedure, refer to the discussion in the *SAS Procedures Guide*.

VAR Statement

VAR *variables* **;**

The VAR statement identifies the variables containing the marker alleles. The VAR statement should contain $2m$ variable names, where m is the number of markers in the data set. Note that alleles for the same marker should be listed consecutively.

Details

Statistical Computations

Frequency Estimates

A marker locus **M** may have a series of alleles M_u, $u = 1, ..., k$. A sample of n individuals may therefore have several different genotypes at the locus, with n_{uv} copies of type M_u/M_v. The number n_u of copies of allele M_u can be found directly by summation: $n_u = 2n_{uu} + \sum_{v \neq u} n_{uv}$. The sample frequencies are written as $\tilde{p}_u = n_u/(2n)$ and $\tilde{P}_{uv} = n_{uv}/n$. The $\tilde{P}_{uv}$'s are unbiased maximum likelihood estimates (MLEs) of the population proportions P_{uv}.

The variance of the sample allele frequency $\tilde{p}_u$ is calculated as

$$\text{Var}(\tilde{p}_u) = \frac{1}{2n}(p_u + P_{uu} - 2p_u^2)$$

and can be estimated by replacing p_u and P_{uu} with their sample values $\tilde{p}_u$ and $\tilde{P}_{uu}$. The variance of the sample genotype frequency $\tilde{P}_{uv}$ is not generally calculated; instead, an MLE of the HWD coefficient D_{uv} for alleles M_u and M_v is calculated as

$$\hat{D}_{uv} = \begin{cases} \tilde{P}_{uv} - \tilde{p}_u\tilde{p}_v, & u = v \\ \tilde{p}_u\tilde{p}_v - \frac{1}{2}\tilde{P}_{uv}, & u \neq v \end{cases}$$

and the MLE's variance is estimated using one of the following formulas, depending on whether the two alleles are the same or different:

$$\begin{aligned} \text{Var}(\hat{D}_{uu}) &= \frac{1}{n}\left[\tilde{p}_u^2(1-\tilde{p}_u)^2 + (1-2\tilde{p}_u)^2\hat{D}_{uu} - \hat{D}_{uu}^2\right] \\ \text{Var}(\hat{D}_{uv}) &= \frac{1}{2n}\Big\{\tilde{p}_u\tilde{p}_v(1-\tilde{p}_u)(1-\tilde{p}_v) + \sum_{w \neq u,v}(\tilde{p}_u^2\hat{D}_{vw} + \tilde{p}_v^2\hat{D}_{uw}) \\ &\quad - \left[(1-\tilde{p}_u-\tilde{p}_v)^2 - 2(\tilde{p}_u-\tilde{p}_v)^2\right]\hat{D}_{uv} + \tilde{p}_u^2\tilde{p}_v^2 - 2\hat{D}_{uv}^2\Big\} \end{aligned}$$

The standard error, the square root of the variance, is reported for the sample allele frequencies and the disequilibrium coefficient estimates. When the BOOTSTRAP=

option of the PROC ALLELE statement is specified, bootstrap confidence intervals are formed by resampling individuals from the data set and are reported for these estimates, with the $100(1-\alpha)$% confidence level given by the ALPHA=α option (or $\alpha = 0.05$ by default).

Measures of Marker Informativeness

Polymorphism Information Content

The polymorphism information content (PIC) measures the probability of differentiating the allele transmitted by a given parent to its child given the marker genotype of father, mother, and child (Botstein et al. 1980). It is computed as

$$\text{PIC} = 1 - \sum_{u=1}^{k} \tilde{p}_u^2 - \sum_{u=1}^{k-1} \sum_{v=u+1}^{k} 2\tilde{p}_u^2 \tilde{p}_v^2$$

Heterozygosity

The heterozygosity, sometimes called the observed heterozygosity, is simply the proportion of heterozygous individuals in the data set and is calculated as

$$\text{Het} = 1 - \sum_{u=1}^{k} \tilde{P}_{uu}$$

Allelic Diversity

The allelic diversity, sometimes called the expected heterozygosity, is the expected proportion of heterozygous individuals in the data set when HWE holds and is calculated as

$$\text{Div} = 1 - \sum_{u=1}^{k} \tilde{p}_u^2$$

Testing for Hardy-Weinberg Equilibrium

Under ideal population conditions, the two alleles an individual receives, one from each parent, are independent so that $P_{uu} = p_u^2$ and $P_{uv} = 2p_u p_v, u \neq v$. The factor of 2 for heterozygotes recognizes the fact that M_u/M_v and M_v/M_u genotypes are generally indistinguishable. This statement about allelic independence within loci is called Hardy-Weinberg equilibrium (HWE). Forces such as selection, mutation, and migration in a population or nonrandom mating can cause departures from HWE. Two methods are used here for testing a marker for HWE, both of which can accommodate any number of alleles. Both methods are testing the hypothesis that $P_{uu} = p_u^2$ and $P_{uv} = 2p_u p_v, u \neq v$ for all $u, v = 1, ..., k$.

Chi-Square Goodness-of-Fit Test

The chi-square goodness-of-fit test can be used to test markers for HWE. The chi-square statistic

$$X_T^2 = \sum_u \frac{(n_{uu} - n\tilde{p}_u^2)^2}{n\tilde{p}_u^2} + \sum_u \sum_{v \neq u} \frac{(n_{uv} - 2n\tilde{p}_u\tilde{p}_v)^2}{2n\tilde{p}_u\tilde{p}_v}$$

has $k(k-1)/2$ degrees of freedom where k is the number of alleles at the marker locus.

Permutation Version of Exact Test

The exact test given by Guo and Thompson (1992) is based on the conditional probability of genotype counts given allelic counts and the hypothesis of allelic independence. The test statistic is

$$T = \frac{n!}{(2n)!} \frac{2^h \prod_u n_u!}{\prod_{u,v} n_{uv}!}$$

where $h = \sum_u \sum_{v \neq u} n_{uv}$ is the number of heterozygous individuals. Significance levels are calculated by the permutation procedure. The $2n$ alleles are randomly permuted the number of times indicated in the EXACT= option to form new sets of n genotypes. The significance level is then calculated as the proportion of times the value of T for each set of permuted data exceeds the value of T for the actual data. You can indicate the random seed used to randomly permute the data in the SEED= option of the PROC ALLELE statement.

Linkage Disequilibrium (LD)

The set of genetic material an individual receives from each parent contains an allele at every locus, and statements can be made about these allelic combinations, or haplotypes. The probability p_{uv} (called the gametic or haplotype frequency) that an individual receives the haplotype $M_u N_v$ for marker loci **M** and **N** can be compared to the product of the probabilities that each allele is received. The difference is the linkage, or gametic, disequilibrium (LD) coefficient D_{uv} for those two alleles: $D_{uv} = p_{uv} - p_u p_v$. There is a general expectation that the amount of linkage disequilibrium is inversely related to the distance between the two loci, but there are many other factors that may affect disequilibrium. There may even be disequilibrium between alleles at loci that are located on different chromosomes. Note that these tests and measures will be calculated only for pairs of markers at most d markers apart, where d is the integer specified in the MAXDIST= option of the PROC ALLELE statement, or 50 by default.

Table 2.1 displays how the HAPLO= option of the PROC ALLELE statement interacts with the linkage disequilibrium calculations. These calculations are discussed in more detail in the following two sections.

Table 2.1. Interaction of HAPLO= option with LD calculations

HAPLO= Option	LD Test Statistic	LD Exact Test	Estimate of Haplotype Freq
GIVEN	$\tilde{D}_{uv}$	Permutes alleles to form new 2-locus haplotypes	Observed freq, $\tilde{p}_{uv}$
EST	$\hat{D}_{uv}$	Not performed	Estimated freq, $\hat{p}_{uv}$
NONE	$\tilde{\Delta}_{uv}$	Permutes genotypes to form new 2-locus genotypes	Composite freq, $\tilde{p}^*_{uv}$

Tests

When haplotypes are known, the HAPLO=GIVEN option should be included in the PROC ALLELE statement so that the linkage disequilibrium can be computed directly by substituting the observed frequencies $\tilde{p}_{uv}$, $\tilde{p}_u$, and $\tilde{p}_v$ into the equation in the preceding section for D_{uv}. This creates the MLE, $\tilde{D}_{uv}$, of the LD coefficient between a pair of alleles at different markers. PROC ALLELE calculates an overall chi-square statistic to test that all of the D_{uv}'s between two markers are zero as follows:

$$X^2_T = \sum_{u=1}^{k} \sum_{v=1}^{l} \frac{(2n)\tilde{D}^2_{uv}}{\tilde{p}_u \tilde{p}_v}$$

which has $(k-1)(l-1)$ degrees of freedom for markers with k and l alleles, respectively.

There is also an exact test available when haplotypes are known. An exact p-value for testing the hypothesis in the preceding paragraph can be calculated by conditioning on the allele counts as with the exact test for HWE. The conditional probability of the haplotype counts is then

$$T = \frac{\prod_u n_u! \prod_v n_v!}{(2n)! \prod_{u,v} n_{uv}!}$$

and the significance level is obtained again by permuting the alleles at one locus to form $2n$ new two-locus haplotypes. You can indicate the number of permutations that are used in the EXACT= option of the PROC ALLELE statement and the random seed used to randomly permute the data in the SEED= option of the PROC ALLELE statement.

When it is requested that haplotype frequencies be estimated with the HAPLO=EST option, D_{uv} is estimated using $\hat{D}_{uv} = \hat{p}_{uv} - \tilde{p}_u \tilde{p}_v$, where $\hat{p}_{uv}$ is the MLE of p_{uv} assuming HWE. The estimate $\hat{p}_{uv}$ is calculated according to the method described by Weir and Cockerham (1979). Again, a chi-square test statistic can be calculated to test that all of the D_{uv}'s between a pair of markers are zero as

$$X^2_T = \sum_{u=1}^{k} \sum_{v=1}^{l} \frac{n\hat{D}^2_{uv}}{\tilde{p}_u \tilde{p}_v}$$

which has $(k-1)(l-1)$ degrees of freedom for markers with k and l alleles, respectively. No exact test is available when haplotype frequencies are estimated.

The HAPLO=NONE option indicates that haplotypes are unknown and $\hat{D}_{uv}$ should not be used in the tests for LD between pairs of markers. Instead of using the estimated haplotype frequencies which assumes HWE, a test can be formed using the composite linkage disequilibrium (CLD) coefficient Δ_{uv} that does not require this assumption and uses only allele and two-locus genotype frequencies. The MLE $\tilde{\Delta}_{uv}$ of Δ_{uv} can be calculated as described by Weir (1979), and a chi-square statistic that tests all Δ_{uv}'s between a pair of markers are zero can be formed as follows:

$$X_T^2 = \sum_{u=1}^{k} \sum_{v=1}^{l} \frac{n\tilde{\Delta}_{uv}^2}{\tilde{p}_u \tilde{p}_v}$$

which has $(k-1)(l-1)$ degrees of freedom for markers with k and l alleles, respectively.

An exact test for CLD is also available, where the conditional probability of the two-locus genotypes given the one-locus genotypes is

$$T = \frac{\prod_{r,s} n_{rs}! \prod_{u,v} n_{uv}!}{n! \prod_{r,s,u,v} n_{rsuv}!}$$

where n_{rsuv} is the count of $M_r M_s N_u N_v$ genotypes, n_{rs} is the count of M_r/M_s genotypes, and n_{uv} is the count of N_u/N_v genotypes. The exact significance level is obtained by permuting the genotypes at one of the loci to create a distribution of the T's (Zaykin, Zhivotovsky, and Weir 1995). Note that this procedure is also testing for nonzero trigenic and quadrigenic disequilibrium terms, so significance may not necessarily imply the presence of CLD.

Measures

PROC ALLELE offers five linkage disequilibrium measures to be calculated for each pair of alleles M_u and N_v located at loci **M** and **N** respectively: the correlation coefficient r, the population attributable risk δ, Lewontin's D', the proportional difference d, and Yule's Q. The five measures are discussed in Devlin and Risch (1995). Since these measures are designed for biallelic markers, the measures are calculated for each allele at locus **M** with each allele at locus **N**, where all other alleles at each loci are combined to represent one allele. Thus for each allele M_u in turn, $\tilde{p}_1$ is used as the frequency of allele M_u, and $\tilde{p}_2$ represents the frequency of "not M_u"; similarly for each N_v in turn, $\tilde{q}_1$ represents the frequency of allele N_v, and $\tilde{q}_2$ the frequency of "not N_v." All measures have the same numerator, $D = p_{11}p_{22} - p_{12}p_{21}$, the LD coefficient, which can be directly estimated using the observed haplotype frequencies $\tilde{p}_{uv}$ when HAPLO=GIVEN, or estimated using the MLEs of the haplotype frequencies $\hat{p}_{uv}$ assuming HWE when HAPLO=EST. The computations for the measures are as follows:

$$r = \frac{D}{(p_1 p_2 q_1 q_2)^{1/2}}$$

$$
\begin{aligned}
\delta &= \frac{D}{q_1 p_{22}} \\
D' &= \frac{D}{D_{\max}}, D_{\max} = \begin{cases} \min(p_1 q_2, q_1 p_2), & D > 0 \\ \min(p_1 q_1, q_2 p_2), & D < 0 \end{cases} \\
d &= \frac{D}{q_1 q_2} \\
Q &= \frac{D}{p_{11} p_{22} + p_{12} p_{21}}
\end{aligned}
$$

with estimates of measures calculated by replacing parameters with their appropriate estimates. Under the default option HAPLO=NONE, the numerator D can be replaced by the CLD coefficient Δ, described in the preceding section, for measures r and D'. This statistic has bounds twice as large as D so the denominator for D' must be multiplied by a factor of 2. However, δ, d, and Q cannot be calculated when HAPLO=NONE.

Missing Values

An individual's genotype for a marker is considered missing if at least one of the alleles at the marker is missing. Any missing genotypes are excluded from all calculations, including the linkage disequilibrium statistics for all pairs that include the marker. However, the individual's nonmissing genotypes at other markers can be used as part of the calculations.

If the BOOTSTRAP= option is specified, any individuals with missing genotypes for all markers are excluded from resampling. All other individuals are included, which could result in different numbers of individuals with nonmissing genotypes for the same marker across different samples.

OUTSTAT= Data Set

The OUTSTAT= data set contains the following variables:

- the BY variables, if any
- Locus1 and Locus2, which contain the pair of markers for which the disequilibrium statistics are calculated
- NIndiv, which contains the number of individuals that have been genotyped at both the markers listed in Locus1 and Locus2 (that is, the number of individuals that have no missing alleles for the two loci)
- Test, which indicates which disequilibrium test is performed, HWE for individual markers (when Locus1 and Locus2 contain the same value) or LD for marker pairs
- ChiSq, which contains the chi-square statistic for testing for disequilibrium. If Locus1 and Locus2 contain the same marker, the test is for HWE within that locus. Otherwise, it is a test for linkage disequilibrium between the two loci.
- DF, which contains the degrees of freedom for the chi-square test

- ProbChi, which contains the p-value for the chi-square test
- ProbEx, which contains the exact p-value for testing the pair of markers in Locus1 and Locus2 for disequilibrium. This variable is included in the OUTSTAT= data set only when the EXACT= parameter in the PROC ALLELE statement is a positive integer and HAPLO=NONE or HAPLO=GIVEN.

Displayed Output

This section describes the displayed output from PROC ALLELE. See the section "ODS Table Names" on page 27 for details about how this output interfaces with the Output Delivery System.

Marker Summary

The "Marker Summary" table lists information on each of the markers, including

- NIndiv, the number of individuals genotyped at the marker
- NAllele, the number of alleles at the marker
- PIC, the polymorphism information content (PIC) measure
- Het, the heterozygosity measure
- Div, the allelic diversity measure

as well as the following columns for the test for HWE:

- ChiSq, the chi-square statistic
- DF, the degrees of freedom for the chi-square test
- ProbChiSq, the p-value for the chi-square test
- ProbExact, the exact p-value for the HWE test (only if the EXACT= option is specified in the PROC ALLELE statement)

Allele Frequencies

The "Allele Frequencies" table lists all the observed alleles for each marker, with an estimate of the allele frequency, the standard error of the frequency, and when the BOOTSTRAP= option is specified, the bootstrap lower and upper limits of the confidence interval for the frequency based on the confidence level determined by the ALPHA= option of the PROC ALLELE statement (0.95 by default).

Genotype Frequencies

The "Genotype Frequencies" table lists all the observed genotypes (denoted by the two alleles separated by a "/") for each marker, with the observed genotype frequency, an estimate of the disequilibrium coefficient D, the standard error of the estimate, and when the BOOTSTRAP= option is specified, the lower and upper limits of the bootstrap confidence interval for D based on the confidence level determined by the ALPHA= option of the PROC ALLELE statement (0.95 by default).

Linkage Disequilibrium Measures

The "Linkage Disequilibrium Measures" table lists the frequency of each haplotype at each marker pair (observed frequency when HAPLO=GIVEN and estimated frequency otherwise), an estimate of the LD coefficient D_{uv}, and whichever linkage disequilibrium measures are included in the PROC ALLELE statement (CORRCOEFF, DELTA, DPRIME, PROPDIFF, and YULESQ). Haplotypes are represented by the allele at the marker locus listed in Locus1 and the allele at the marker locus listed in Locus2 separated by a "-." Note that this table can be quite large when there are many markers or markers with many alleles. For a data set with m markers, each having k_i alleles, $i = 1, ..., m$, the number of rows in the table is $\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} k_i k_j$. The MAXDIST= option of the PROC ALLELE statement can be used to keep this table to a manageable size.

ODS Table Names

PROC ALLELE assigns a name to each table it creates, and you must use this name to reference the table when using the Output Delivery System (ODS). These names are listed in the following table.

Table 2.2. ODS Tables Created by the ALLELE Procedure

ODS Table Name	Description	PROC ALLELE option
MarkerSumm	Marker summary	default
AlleleFreq	Allele frequencies	default
GenotypeFreq	Genotype frequencies	default
LDMeasures	Linkage disequilibrium measures	CORRCOEFF, DELTA, DPRIME, PROPDIFF, or YULESQ

Examples

Example 2.1. Using the NDATA= Option with Microsatellites

The following is a subset of data from GAW12 (Wijsman et al. 2001) and contains 17 individuals' genotypes at 14 microsatellite markers.

```
data gaw;
   input id m1-m28;
   datalines;
 1 11 14  6  8  2  5  9  4  6  1  9  9  9  7  3  5 10  1  4  6  5  9  1  1  3  5  6
 2  2 12  1  4  6  6  3  3  2  1 11 11  4 11  2  2 13 11  2  1  9  9  1  5  6  1  2
 3  2 10  4  8  4  9  2  7  7  1  9  2  7 10  2  2  7  7  6  8  9  4  5  1  7  2  6
 4  5 14  7  3  9 13  4  2  2  4 11  5  4  7  4  5  7  6  8  2  9  9  1  6  4  1  8
 5 12 12  3  8  6  2  1  7  3  5  6 11  6  9  5  2 13 16  7  1  9  4  1  1  7  1  1
 6  4  7  7  8  7 12  4  2  6  5  5 11  5 11  2  4 15 11  1  1  9  2  6  5  7  6  1
 7  2 10  6  8  7  1  2  3  6  2  5  8  5  6  5  6 13 10  1  8  9  3  1  6  7  7  2
 8  2 11  6  2  7  1  2  3  6  6 10 11 11  6  4  2 11 11  4  5 11  2  3  2  1  4  1
 9  2  7  1  1  3  1  5  7  2  5  5 11 11 11  2  6 11  2  1  6  4  9  5  5  4  2  5
```

```
10 11 12  2  4 13  3  1  2  4  9  5 10  7  5  4  4  1  6  8  1  6 10  1  1  2  5  1  1
11 11  2  7  8  1  5  4  6  4  7  5 11 11  6  5  4 16 13  7  4  5  6  6  1  1  4  1  1
12  2 12  6  8  2  7  3  2  7  5  2  8  9  6  2  4  7 16  7  1 10  9  5  1  1  4  9  1
13 13 14  8  3 12 13  7  4  3  2  6 10  9  5  4  4  2 14  8  8  3  6  5  1  1  6  6  2
14  7 10  6  5 10 13  8  3  5  5  9  9 11  6  5  4 13 14  1  1  6  9  2  1  5  3  1  2
15 10 11  4  3  9  7  6  3  4  6 10  1  7  9  2  2  2 14  6  1  9  2  1  1  6  7  5  2
16  2  5  2  7  7  2  2  9  2  2  2  6  9  5  2  2  7  1  1  2  6  2  1  1  1  1  9  6
17 11  4  4  4  9  1  7  8  5  3  5  1 11  5  6  5  2 12  1  5  9  9  1  5  7  7  6  1
;
```

Note that you can input the same data directly using the statement:

```
infile 'Genmrk22.1' delimiter="/ ";
```

in place of the DATALINES statement.

The actual names of the markers can be used, by creating a data set with the variable NAME containing these names.

```
data map;
   input name $ location;
   datalines;
D22G001   0.50
D22G002   0.79
D22G003   0.88
D22G004   1.02
D22G005   1.24
D22G006   2.20
D22G007   4.27
D22G008   5.85
D22G009   6.70
D22G010   9.36
D22G011  10.87
D22G012  11.67
D22G013  12.66
D22G014  15.89
;
```

Now an analysis using PROC ALLELE can be performed as follows:

```
proc allele data=gaw ndata=map nofreq exact=10000 seed=456;
   var m1-m28;
run;
```

This analysis produces summary statistics of the 14 markers and is using 10,000 permutations to get an exact p-value for the HWE test. The allele and genotype frequency output tables are suppressed with the NOFREQ option.

The results from the analysis are as follows. Note the names of the markers that are used.

Output 2.1.1. Summary of Microsatellites for the ALLELE Procedure

The ALLELE Procedure

Marker Summary

Locus	Number of Indiv	Number of Alleles	PIC	Hetero-zygosity	Allelic Diversity
D22G001	17	9	0.8384	0.9412	0.8547
D22G002	17	8	0.8296	0.8824	0.8478
D22G003	17	11	0.8749	0.9412	0.8858
D22G004	17	9	0.8259	0.9412	0.8443
D22G005	17	8	0.8272	0.8235	0.8460
D22G006	17	8	0.8257	0.8235	0.8443
D22G007	17	7	0.8012	0.9412	0.8253
D22G008	17	5	0.6665	0.6471	0.7163
D22G009	17	11	0.8788	0.8824	0.8893
D22G010	17	7	0.7572	0.8235	0.7820
D22G011	17	8	0.7274	0.8235	0.7509
D22G012	17	5	0.5661	0.6471	0.6142
D22G013	17	7	0.7965	0.8235	0.8201
D22G014	17	6	0.7507	0.8824	0.7837

Marker Summary

	Test for HWE			
Locus	Chi-Square	DF	Pr > ChiSq	Prob Exact
D22G001	32.5172	36	0.6350	0.8581
D22G002	28.5222	28	0.4370	0.3868
D22G003	48.2139	55	0.7295	0.7050
D22G004	24.9692	36	0.9166	0.8361
D22G005	20.9416	28	0.8278	0.9413
D22G006	32.0018	28	0.2744	0.1102
D22G007	19.7625	21	0.5363	0.5745
D22G008	11.4619	10	0.3227	0.2525
D22G009	52.1333	55	0.5849	0.3866
D22G010	14.7227	21	0.8366	0.8624
D22G011	19.0400	28	0.8969	0.8898
D22G012	17.3473	10	0.0670	0.5122
D22G013	38.8062	21	0.0104	0.0390
D22G014	17.2802	15	0.3024	0.4651

Example 2.2. Computing Linkage Disequilibrium Measures for SNP Data

The following data set contains 44 individuals' genotypes at five SNPs.

```
data snps;
   input s1-s10;
   datalines;
2 2 2 1 2 1 1 1 2 2
2 2 2 2 2 1 1 1 2 2
2 2 2 2 2 1 2 1 2 2
2 2 2 2 . . 1 1 2 2
2 2 2 2 1 2 1 2 2 2
2 2 2 2 . . 2 1 2 2
```

```
2 2 2 2 2 1 2 1 2 2
2 2 2 2 . . 2 1 2 2
2 2 2 2 1 1 1 1 2 2
2 2 1 1 2 2 2 1 2 2
2 2 2 1 2 2 2 1 2 2
2 2 2 2 1 1 1 1 2 2
2 2 2 1 2 2 2 2 2 2
2 2 2 2 2 2 1 1 2 2
2 2 2 2 2 1 2 1 2 2
2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 1 2 2
2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 1 1 1 2 2
2 2 2 2 1 1 2 1 2 2
2 2 2 2 2 1 1 1 2 2
2 2 2 2 2 1 2 2 2 2
2 2 2 2 2 1 2 1 2 2
2 2 2 2 2 1 2 1 2 2
2 2 2 2 2 1 2 2 2 2
2 2 2 2 2 1 2 2 2 2
2 2 2 2 2 2 1 1 2 2
2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 1 1 2 2
2 2 2 2 2 2 1 1 2 2
2 2 2 2 2 1 2 1 2 2
2 2 2 2 2 1 2 1 2 2
2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 1 2 2
2 2 2 2 2 1 2 1 2 2
2 2 2 2 2 2 . . 2 2
2 2 2 1 2 2 2 1 2 2
2 2 2 2 2 2 2 1 2 2
2 2 2 2 2 1 1 1 2 2
2 2 2 2 2 2 1 1 2 2
2 2 2 2 2 1 2 1 2 2
2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 1 2 2
2 2 2 2 2 2 2 1 2 2
;
```

Now an analysis using PROC ALLELE can be performed as follows:

```
proc allele data=snps prefix=SNP nofreq haplo=est corrcoeff dprime yulesq;
   var s1-s10;
run;
```

This analysis produces summary statistics of the five SNPs as well as the Linkage Disequilibrium Measures table, which contains estimated two-locus haplotype frequencies and disequilibrium coefficients, and the linkage disequilibrium measures r, D', and Q. The allele and genotype frequency output tables are suppressed with the NOFREQ option.

The results from the analysis are as follows. Note the names of the markers that are used.

Output 2.2.1. Summary of SNPs for the ALLELE Procedure

The ALLELE Procedure

Marker Summary

Locus	Number of Indiv	Number of Alleles	PIC	Hetero-zygosity	Allelic Diversity	Test for HWE: Chi-Square	DF	Pr > ChiSq
SNP1	44	1	0.0000	0.0000	0.0000	0.0000	0	.
SNP2	44	2	0.1190	0.0909	0.1271	3.5627	1	0.0591
SNP3	41	2	0.3283	0.4390	0.4140	0.1493	1	0.6992
SNP4	43	2	0.3728	0.4884	0.4957	0.0093	1	0.9231
SNP5	44	1	0.0000	0.0000	0.0000	0.0000	0	.

There are two SNPs that have only one allele appearing in the data.

Output 2.2.2. Linkage Disequilibrium Measures for SNPs Using the ALLELE Procedure

Linkage Disequilibrium Measures

Locus1	Locus2	Haplotype	Frequency	LD Coeff	Corr Coeff	Lewontin's D'	Yule's Q
SNP1	SNP2	2-1	0.0682	-0.000	.	.	.
SNP1	SNP2	2-2	0.9318	-0.000	.	.	.
SNP1	SNP3	2-1	0.2927	-0.000	.	.	.
SNP1	SNP3	2-2	0.7073	-0.000	.	.	.
SNP1	SNP4	2-1	0.5465	-0.000	.	.	.
SNP1	SNP4	2-2	0.4535	-0.000	.	.	.
SNP1	SNP5	2-2	1.0000	0.000	.	.	.
SNP2	SNP3	1-1	0.0000	-0.021	-0.181	-1.000	-1.000
SNP2	SNP3	1-2	0.0732	0.021	0.181	1.000	1.000
SNP2	SNP3	2-1	0.2927	0.021	0.181	1.000	1.000
SNP2	SNP3	2-2	0.6341	-0.021	-0.181	-1.000	-1.000
SNP2	SNP4	1-1	0.0331	-0.005	-0.040	-0.132	-0.155
SNP2	SNP4	1-2	0.0367	0.005	0.040	0.132	0.155
SNP2	SNP4	2-1	0.5134	0.005	0.040	0.132	0.155
SNP2	SNP4	2-2	0.4168	-0.005	-0.040	-0.132	-0.155
SNP2	SNP5	1-2	0.0682	-0.000	.	.	.
SNP2	SNP5	2-2	0.9318	-0.000	.	.	.
SNP3	SNP4	1-1	0.2221	0.061	0.266	0.438	0.553
SNP3	SNP4	1-2	0.0779	-0.061	-0.266	-0.438	-0.553
SNP3	SNP4	2-1	0.3154	-0.061	-0.266	-0.438	-0.553
SNP3	SNP4	2-2	0.3846	0.061	0.266	0.438	0.553
SNP3	SNP5	1-2	0.2927	-0.000	.	.	.
SNP3	SNP5	2-2	0.7073	-0.000	.	.	.
SNP4	SNP5	1-2	0.5465	-0.000	.	.	.
SNP4	SNP5	2-2	0.4535	-0.000	.	.	.

In the preceding table, the values for the linkage disequilibrium measures are missing for several haplotypes; this occurs when there is only one allele at one of the markers contained in the haplotype, and thus the denominators for these measures are zero.

Also note that when the markers are biallelic, the gametic disequilibria have the same absolute values for all four possible haplotypes.

References

Botstein, D., White, R.L., Skolnick, M., and Davis, R.W. (1980), "Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms," *American Journal of Human Genetics*, 32, 314–331.

Devlin, B. and Risch, N. (1995), "A Comparison of Linkage Disequilibrium Measures for Fine-Scale Mapping," *Genomics*, 29, 311–322.

Guo, S.W. and Thompson, E.A. (1992), "Performing the Exact Test of Hardy-Weinberg Proportion for Multiple Alleles," *Biometrics*, 48, 361–372.

Weir, B.S. (1979), "Inferences about Linkage Disequilibrium," *Biometrics*, 35, 235–254.

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

Weir, B.S. and Cockerham, C.C. (1979), "Estimation of Linkage Disequilibrium in Randomly Mating Populations," *Heredity*, 42, 105–111.

Wijsman, E.M., Almasy, L., Amos, C.I., Borecki, I., Falk, C.T., King, T.M., Martinez, M.M., Meyers, D., Neuman, R., Olson, J.M., Rich, S., Spence, M.A., Thomas, D.C., Vieland, V.J., Witte, J.S., and MacCluer, J.W. (2001), "Analysis of Complex Genetic Traits: Applications to Asthma and Simulated Data," *Genetic Epidemiology*, 21, S1–S853.

Zaykin, D., Zhivotovsky, L., and Weir, B.S. (1995), "Exact Tests for Association between Alleles at Arbitrary Numbers of Loci," *Genetica*, 96, 169–178.

Chapter 3
The CASECONTROL Procedure

Chapter Contents

Chapter 3
The CASECONTROL Procedure

Overview

Marker information can be used to help locate the genes that affect susceptibility to a disease. The CASECONTROL procedure is designed for the interpretation of marker data when random samples are available from the populations of unrelated individuals who are either affected or unaffected by the disease. Several tests are available in PROC CASECONTROL that compare marker allele and/or genotype frequencies in the two populations, with frequency differences indicating an association of the marker with the disease. Although such an association may point to the proximity of the marker and disease genes in the genome, it may also reflect population structure, so care is needed in interpreting the results; association does not necessarily imply linkage.

The three chi-square tests available for testing case-control genotypic data are the genotype case-control test, which tests for dominant allele effects on the disease penetrance, and the allele case-control test and linear trend test, which test for additive allele effects on the disease penetrance. Since the allele case-control test requires the assumption of Hardy-Weinberg equilibrium (HWE), it may be desirable to run the ALLELE procedure on the data to perform the HWE test on each marker (see Chapter 2, "The ALLELE Procedure," for more information) prior to applying PROC CASECONTROL.

Getting Started

Example

Here are some sample SNP data on which the three case-control tests can be performed using PROC CASECONTROL:

```
data cc;
   input affected $ m1-m16;
   datalines;
 N  1 1 2 2 2 2 2 1 2 1 2 2 1 1 2 2
 N  1 1 1 1 2 2 1 1 2 1 2 1 1 1 1 1
 N  2 1 1 1 2 1 1 1 2 2 1 1 1 1 1 1
 N  2 2 2 1 2 2 1 1 2 2 2 1 1 1 2 2
 N  1 1 1 1 2 2 2 1 1 1 1 1 2 1 . .
 N  2 1 1 1 2 1 1 1 2 1 2 1 1 1 2 1
 N  1 1 1 1 2 2 1 1 2 2 2 2 2 1 2 2
 N  2 2 1 1 2 1 2 1 2 2 2 1 1 1 2 1
 N  2 1 1 1 2 2 2 1 2 1 . . 1 1 2 1
 N  2 1 1 1 2 1 1 1 2 2 1 1 1 1 1 1
 N  2 1 2 2 . . 1 1 2 1 1 1 1 1 1 1
```

```
N  2 2 . . 2 1 1 1 2 1 2 1 1 1 2 1
N  2 1 . . 2 2 1 1 2 2 1 1 1 1 2 1
N  2 1 . . 2 2 1 1 2 1 . . 2 1 1 1
N  2 2 . . 2 2 1 1 . . 2 1 1 1 2 1
N  1 1 . . 2 2 1 1 1 1 2 1 1 1 2 1
N  1 1 . . 2 2 1 1 1 1 . . 1 1 2 1
N  2 1 . . 2 2 1 1 1 1 . . 2 1 2 1
A  2 1 2 1 2 1 1 1 1 1 2 1 . . 2 1
A  2 1 2 1 2 2 1 1 2 1 1 1 . . 1 1
A  2 2 2 1 2 2 1 1 2 2 . . . . 2 1
A  2 1 2 2 2 1 1 1 2 1 2 1 . . 2 2
A  . . 2 2 2 1 . . 1 1 2 2 . . 2 1
A  1 1 1 1 2 1 1 1 2 1 1 1 . . 2 2
A  2 1 1 1 2 2 1 1 1 1 2 1 . . 2 1
A  2 1 2 2 2 2 1 1 2 2 . . . . 2 2
A  2 1 1 1 2 2 1 1 2 1 2 1 . . 1 1
A  2 1 2 2 2 1 1 1 2 1 2 1 . . 2 2
A  1 1 1 1 2 2 1 1 2 1 2 1 . . 2 2
A  2 1 2 1 2 1 1 1 2 1 2 2 . . 2 1
A  2 2 2 2 1 1 1 1 2 1 2 1 . . 2 2
A  1 1 1 1 2 1 . . 2 1 2 2 . . 2 2
A  1 1 2 1 2 1 1 1 2 1 2 1 . . 2 2
A  2 2 1 1 2 2 1 1 2 1 1 1 . . 2 1
;
```

The following SAS code can be used to perform the analysis:

```
proc casecontrol data=cc prefix=Marker;
   var m1-m16;
   trait affected;
run;

proc print;
run;
```

All three case-control tests are performed by default. The output data set created by default appears as follows:

Obs	Locus	ChiSq Genotype	ChiSq Allele	Chi Sq Trend	df Genotype	df Allele	df Trend	Prob Genotype	Prob Allele	Prob Trend
1	Marker1	0.272	0.033	0.032	2	1	1	0.873	0.857	0.858
2	Marker2	3.430	3.260	2.140	2	1	1	0.180	0.071	0.144
3	Marker3	2.981	2.569	2.925	2	1	1	0.225	0.109	0.087
4	Marker4	3.556	3.319	3.556	2	1	1	0.169	0.069	0.059
5	Marker5	3.004	0.535	0.590	2	1	1	0.223	0.464	0.443
6	Marker6	0.767	0.650	0.710	2	1	1	0.682	0.420	0.399
7	Marker7	0.000	0.000	0.000	0	0	0	.	.	.
8	Marker8	4.132	4.061	3.769	2	1	1	0.127	0.044	0.052

Figure 3.1. Statistics for Case-Control Tests

Figure 3.1 displays the statistics for the three tests. The genotype case-control statistic has more degrees of freedom than the other two because it is testing for both dominance genotypic effects and additive effects, while the other statistics are testing for the significant additive effects alone. The p-values for Marker7 are missing because the genotypes of all the affected individuals are missing at that marker.

Syntax

The following statements are available in PROC CASECONTROL.

PROC CASECONTROL < *options* > **;**
BY *variables* **;**
TRAIT *variable* **;**
VAR *variables* **;**

Items within angle brackets (< >) are optional, and statements following the PROC CASECONTROL statement can appear in any order. The TRAIT and VAR statements are required. The syntax of each statement is described in the following section in alphabetical order after the description of the PROC CASECONTROL statement.

PROC CASECONTROL Statement

PROC CASECONTROL < *options* > **;**

You can specify the following options in the PROC CASECONTROL statement.

ALLELE
requests that the allele case-control test be performed. If none of the three test options (ALLELE, GENOTYPE, or TREND) are specified, then all three tests are performed by default.

DATA=*SAS-data-set*
names the input SAS data set to be used by PROC CASECONTROL. The default is to use the most recently created data set.

GENOTYPE
requests that the genotype case-control test be performed. If none of the three test options (ALLELE, GENOTYPE, or TREND) are specified, then all three tests are performed by default.

NDATA=*SAS-data-set*
names the input SAS data set containing names, or identifiers, for the markers used in the output. There must be a NAME variable in this data set, which should contain the same number of rows as there are markers in the input data set specified in the DATA= option. When there are fewer rows than there are markers, markers without a name are named using the PREFIX= option. Likewise, if there is no NDATA= data set specified, the PREFIX= option is used.

OUTSTAT=*SAS-data-set*
names the output SAS data set containing the chi-square statistics, degrees of freedom, and p-values for the tests performed. When this option is omitted, an output data set is created by default and named according to the DATA*n* convention.

PREFIX=*prefix*
specifies a prefix to use in constructing names for marker variables in all output. For example, if PREFIX=VAR, the names of the variables are VAR1, VAR2, ..., VAR*n*. Note that this option is ignored when the NDATA= option is specified, unless there are fewer names in the NDATA data set than there are markers. If this option is omitted, PREFIX=M is the default.

TREND
requests that the linear trend test be performed. If none of the three test options (ALLELE, GENOTYPE, or TREND) are specified, then all three tests are performed by default.

VIF
specifies that the variance inflation factor λ should be applied to the trend chi-square statistic for genomic control. This adjustment is applied only when the trend test is performed and all markers in the VAR statement are biallelic.

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC CASECONTROL to obtain separate analyses on observations in groups defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables. The *variables* are one or more variables in the input data set.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data using the SORT procedure with a similar BY statement.
- Specify the BY statement option NOTSORTED or DESCENDING in the BY statement for the CASECONTROL procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables using the DATASETS procedure (in Base SAS software).

For more information on the BY statement, refer to the discussion in *SAS Language Reference: Concepts*. For more information on the DATASETS procedure, refer to the discussion in the *SAS Procedures Guide*.

TRAIT Statement

TRAIT *variable* **;**

The TRAIT statement identifies a binary variable indicating which individuals are cases and which are controls or a binary variable representing a dichotomous trait. This variable can be character or numeric, but must have only two nonmissing levels.

VAR Statement

VAR *variables* **;**

The VAR statement identifies the variables containing the marker alleles. The VAR statement should contain $2m$ variable names, where m is the number of markers in the data set. Note that alleles for the same marker should be listed consecutively.

Details

Statistical Computations

Biallelic Markers

PROC CASECONTROL offers three statistics to test for an association between a biallelic marker and a binary variable, typically affection status of a particular disease. Table 3.1 displays the quantities that are used for the three case-control tests for biallelic markers (Sasieni 1997).

Table 3.1. Genotype Distribution for Case-Control Sample

	Number of M_1 alleles			
	0	1	2	Total
Case	r_0	r_1	r_2	R
Control	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

The three statistical methods for testing a marker for association with a disease locus are Armitage's trend test (1955), the allele case-control test, and the genotype case-control test. The trend test and allele case-control test are most useful when there is an additive allele effect on the disease susceptibility. When Hardy-Weinberg equilibrium (HWE) holds in the combined sample of cases and controls, these statistics are approximately equal and have an asymptotic χ^2_1 distribution. However, if the assumption of HWE in the combined sample is violated, then the variance for the allele case-control statistic is incorrect; only the trend test remains valid under this violation. The statistics for the trend and allele case-control test, respectively, are given by Sasieni (1997) as

$$X^2_T = \frac{N[N(r_1+2r_2)-R(n_1+2n_2)]^2}{R(N-R)[N(n_1+2n_2)-(n_1+2n_2)^2]}$$

$$X^2_A = \frac{2N[2N(r_1+2r_2)-2R(n_1+2n_2)]^2}{(2R)2(N-R)[2N(n_1+2n_2)-(n_1+2n_2)^2]}$$

Devlin and Roeder (1999) describe a genomic control method that adjusts the trend test statistic for correlation between alleles from members of the same subpopulation. Assuming the variance inflation factor λ is constant across the genome, it can be estimated by $\hat{\lambda} = [\text{median}(X_1, ..., X_m)/0.675]^2$, where $X_i = X_T$ for the ith marker, $i = 1, ..., m$. The adjusted trend statistic, $X^2_{T_a} = X^2_T/\hat{\lambda}$, is approximately distributed as χ^2_1. This variance correction is made when the VIF option is specified in the PROC statement and all markers are biallelic.

If dominance effects of alleles are also suspected to contribute to disease susceptibility, the genotype case-control test can be used. The standard 2×3 contingency table analysis is used to form the χ^2_2 statistic for the genotype case-control test as

$$X^2_G = \sum_{i=0}^{2} \left[\frac{(Nr_i - Rn_i)^2}{NRn_i} + \frac{(Ns_i - Sn_i)^2}{NSn_i} \right]$$

which tests for both additive and dominance (non-additive) allelic effects (Nielsen and Weir 1999). Note that this test can be broken into two 1 df chi-square tests: the trend test mentioned above, and a test for dominance effects alone. This test can be performed by taking the difference $X^2_G - X^2_T = X^2_D$ to form a new test statistic asymptotically distributed as χ^2_1.

Multiallelic Markers

When there are multiple alleles of interest at a marker, the same three tests can be performed, except that Devlin and Roeder's genomic control adjustment is not applied when there are any markers with more than two alleles. To construct the test statistic for the multiallelic trend test for a marker with k alleles (Slager and Schaid 2001), the $p \times (k-1)$ matrix $\mathbf{X}$ is created such that each element X_{iu} represents the number of times the M_u allele appears in the ith genotype, $i = 1, ..., p$ and $u = 1, ..., k-1$, where $p = k(k+1)/2$, the number of possible genotypes. Vectors $\mathbf{r}$ and $\mathbf{s}$ of length p contain the genotype counts for the cases and controls respectively, and $\phi = R/N$, the proportion of cases in the sample. The multiallelic trend test statistic can then be expressed as $\mathbf{U}'[\text{Var}(\mathbf{U})]^{-1}\mathbf{U}$, where the vector $\mathbf{U} = \mathbf{X}'[(1-\phi)\mathbf{r} - \phi\mathbf{s}]$. Var($\mathbf{U}$) is calculated under the assumption of independent (or unrelated) subjects in the sample using Var($\mathbf{r}$) and Var($\mathbf{s}$). These matrices contain elements $\sigma_{ii} = Rn_i(N - n_i)/N^2$ and $\sigma_{ij} = -Rn_in_j/N^2$, where $i, j = 1, ..., p$ (the R is replaced by S for Var($\mathbf{s}$)). This statistic has an asymptotic χ^2_{k-1} distribution.

Another way to test for additive allele effects at the disease or trait locus is the allele case-control test, executed using a contingency table analysis similar to the genotype case-control test described in the preceding section, assuming HWE (Nielsen and Weir 1999). For a marker with k alleles, a 2×k contingency table is formed with one row for cases, one for controls, and a column for each allele. The χ^2_{k-1} statistic is formed by summing $(O - E)^2/E$ over all cells in the table, where O is the observed count for the cell and E is the expected count, the cell's column total multiplied by R/N (or S/N) for a cell in the case (or control) row.

The genotypic case-control test statistic is calculated in a similar manner, with columns now representing the $p = k(k+1)/2$ genotype classes instead of alle-

les. Significance of this test statistic using the χ^2_{p-1} distribution indicates dominance and/or additive allelic effects on the disease or trait (Nielsen and Weir 1999).

Missing Values

An individual's genotype for a marker is considered missing if at least one of the alleles at the marker is missing. Any missing genotypes are excluded from all calculations. However, the individual's nonmissing genotypes at other loci can be used as part of the calculations. If an individual has a missing trait value, then that individual is excluded from all calculations.

OUTSTAT= Data Set

The output data set specified in the OUTSTAT= option of the PROC CASECONTROL statement contains the following variables:

- the BY variables, if any
- Locus
- the chi-square statistic for each test performed: ChiSqAllele, ChiSqGenotype, and ChiSqTrend
- the degrees of freedom for each test performed: dfAllele, dfGenotype, and dfTrend
- the p-value for each test performed: ProbAllele, ProbGenotype, and ProbTrend

Example

Example 3.1. Performing Case-Control Tests on Multiallelic Markers

The following data are taken from GAW9 (Hodge 1995). A sample of 60 founders was taken from 200 nuclear families, 30 affected with a disease and 30 unaffected. Each founder was genotyped at two marker loci.

```
data founders;
   input id disease a1-a4 @@;
   datalines;
4   1 6 4 3 7  17  2 4 7 2 7
39  2 6 8 7 7  41  2 4 4 4 7
46  1 8 4 1 5  50  2 4 2 3 7
54  2 4 8 7 6  56  2 7 4 7 7
62  2 4 1 7 3  69  2 6 8 2 7
79  1 6 6 8 7  80  2 6 4 7 3
83  2 8 4 2 7  85  1 5 6 6 2
95  1 3 2 3 7  101 1 4 6 7 7
106 1 2 1 7 2  107 1 1 2 7 7
115 2 4 2 7 5  116 1 4 1 7 3
```

```
120 2 1 6 2 7  123 2 4 4 7 2
130 1 5 2 3 7  133 1 8 6 3 6
134 1 8 4 2 2  139 2 6 4 7 6
142 2 3 6 7 7  151 1 4 6 4 3
152 1 6 7 6 7  153 1 5 1 7 6
154 1 4 6 6 6  168 1 1 4 3 7
178 2 4 1 7 1  187 1 1 8 1 2
189 2 6 4 5 7  190 2 4 4 3 7
195 2 4 4 7 2  207 2 1 6 7 7
216 1 7 4 1 5  222 2 4 2 7 3
225 2 8 7 7 6  234 1 6 4 2 2
244 1 4 4 7 6  249 2 6 8 7 2
263 1 8 2 3 7  267 2 2 2 2 7
276 2 1 6 7 1  284 2 4 8 2 2
286 1 8 8 2 1  289 1 2 6 6 3
290 1 2 4 5 7  294 2 1 8 6 7
297 2 5 4 7 6  313 1 1 7 7 2
337 1 2 6 7 6  366 2 2 2 7 7
368 2 3 1 7 2  381 1 6 4 5 3
384 1 6 2 2 7  396 1 4 5 7 2
;
```

The multiallelic versions of the association tests are performed since each marker has more than two alleles. The following code invokes the three case-control tests to find out whether there is a significant association between either of the markers and disease status. Note that the same output could be produced by omitting the three tests, ALLELE, GENOTYPE, and TREND, from the PROC CASECONTROL statement.

```
proc casecontrol data=founders genotype allele trend;
   trait disease;
   var a1-a4;
run;

proc print noobs;
run;
```

An output data set is created by default, and the output from the PRINT procedure is displayed in Output 3.1.1.

Output 3.1.1. Output Data Set from PROC CASECONTROL for Multiallelic Markers

Locus	ChiSq Genotype	ChiSq Allele	ChiSq Trend	df Genotype	df Allele	df Trend	Prob Genotype	Prob Allele	Prob Trend
M1	27.333	4.441	5.039	35	7	7	0.8191	0.7278	0.6552
M2	18.077	8.772	13.244	35	7	7	0.9920	0.2694	0.0664

This analysis finds no significant association between disease status and either of the markers. Suppose, however, that allele 7 of the second marker had been identified

by previous studies as an allele of interest for this particular disease, and thus there is concern that its effect is swamped by the other seven alleles. The data set can be modified so that the second marker is considered a biallelic marker with alleles 7 and "not 7."

```
data marker2;
   set founders;
   if a3 ne 7 then a3=1;
   if a4 ne 7 then a4=1;
   keep id a3 a4 disease;
```

Now all three tests can be performed on the marker in the new data set.

```
proc casecontrol data=marker2;
   trait disease;
   var a3 a4;
run;

proc print noobs;
run;
```

PROC CASECONTROL performs all three tests by default since none were specified. The output data set for this analysis is displayed in Output 3.1.2.

Output 3.1.2. Output Data Set from PROC CASECONTROL for a Biallelic Marker

Locus	ChiSq Genotype	ChiSq Allele	ChiSq Trend	df Genotype	df Allele	df Trend	Prob Genotype	Prob Allele	Prob Trend
M1	12.193	6.599	10.103	2	1	1	0.0023	0.0102	0.0015

With just the single allele of interest, there is now a significant association (using a significance level of $\alpha = 0.05$) according to all three case-control tests between the marker (specifically, allele 7) and disease status. Note that the allele and trend tests, both of which are testing for additive allele effects, produce quite different p-values, which could be an indication that HWE does not hold for allele 7. This is in fact the case, which can be checked by running the ALLELE procedure on data set marker2 to test for HWE (see Chapter 2, "The ALLELE Procedure," for more information). The excess of heterozygotes forces X_A^2 to be smaller than X_T^2, and only X_T^2 remains a valid chi-square statistic under the HWE violation.

References

Armitage, P. (1955), "Tests for Linear Trends in Proportions and Frequencies," *Biometrics*, 11, 375–386.

Devlin, B. and Roeder, K. (1999), "Genomic Control for Association Studies," *Biometrics*, 55, 997–1004.

Hodge, S.E. (1995), "An Oligogenic Disease Displaying Weak Marker Associations: A Summary of Contributions to Problem 1 of GAW9," *Genetic Epidemiology*, 12, 545–554.

Nielsen, D.M. and Weir, B.S. (1999), "A Classical Setting for Associations between Markers and Loci Affecting Quantitative Traits," *Genetic Research*, 74, 271–277.

Sasieni, P.D. (1997), "From Genotypes to Genes: Doubling the Sample Size," *Biometrics*, 53, 1253–1261.

Slager, S.L. and Schaid, D.J. (2001), "Evaluation of Candidate Genes in Case-Control Studies: A Statistical Method to Account for Related Subjects," *American Journal of Human Genetics*, 68, 1457–1462.

Chapter 4
The FAMILY Procedure

Chapter Contents

Chapter 4
The FAMILY Procedure

Overview

Family genotype data, though more difficult to collect, often provide a more effective way of testing markers for association with disease status than case-control data. Case-control data may uncover significant associations between markers and a disease that could be caused by factors other than linkage, such as population structure. Analyzing family data using the FAMILY procedure ensures that any significant associations found between a marker and disease status are due to linkage between the marker and disease locus. This is accomplished by using the transmission/disequilibrium test (TDT) and several variations of it that can accommodate different types of family data. One type of family consists of parents, at least one heterozygous, and an affected child who have all been genotyped. This family structure is suitable for the original TDT. Families containing at least one affected and one unaffected sibling from a sibship that have both been genotyped can be analyzed using the sibling tests: the sib TDT (S-TDT) or the nonparametric sibling disequilibrium test (SDT). Both types of families can be jointly analyzed using the combined versions of the S-TDT and SDT and the reconstruction-combined TDT (RC-TDT). The RC-TDT can additionally accommodate families with no unaffected children and missing parental genotypes in certain situations.

Getting Started

Example

The following example demonstrates how you can use PROC FAMILY to perform one of several family-based tests, the TDT. You have collected the following family genotypic data that you input into a SAS data set:

```
data example;
   input ped indiv father mother disease (a1-a4)($);
   datalines;
 1   1  0  0 1 a b a c
 1   2  0  0 1 c c a d
 1 101  1  2 1 a c a d
 1 102  1  2 1 b c a d
 1 103  1  2 1 a c c a
 1 104  1  2 2 b c a a
 2   3  0  0 1 e e f g
 2   4  0  0 1 d e g a
 2 105  3  4 1 e d f a
 2 106  3  4 2 e e g a
 3   5  0  0 1 d a a c
```

```
 3   6  0  0 1 e e c a
 3 107  5  6 2 a e a a
 4   7  0  0 1 f b a g
 4   8  0  0 1 c e h g
 4 108  7  8 2 b e a g
 4 109  7  8 1 f c g g
 4 110  7  8 1 b c a g
 4 111  7  8 1 b c a h
 5   9  0  0 1 a f d c
 5  10  0  0 1 h d c h
 5 112  9 10 2 a d d c
 5 113  9 10 1 f d d c
 6  11  0  0 1 b e c g
 6  12  0  0 1 d f a g
 6 114 11 12 2 b f c a
 6 115 11 12 1 b d g a
 7  13  0  0 1 e d c c
 7  14  0  0 1 e h d a
 7 116 13 14 1 e h c a
 7 117 13 14 2 d e c a
 7 118 13 14 1 d h c d
 7 119 13 14 1 d h c d
;
```

The first column of the data set contains the pedigree ID, followed by an individual ID, and the two parental IDs. The fifth column is a variable representing affection status of a disease. The last four columns of this data set contain the two alleles at each of two markers for each individual. Since there are no missing parental genotypes in this data set, the TDT is a reasonable test to perform in order to determine if either of the two markers is significantly linked to the disease locus whose location you are trying to pinpoint. Furthermore, close inspection of the data reveals that there is only one affected child (which corresponds to a value of "2" for the disease affection variable) per each family. Thus, the TDT is also a valid test for association with the disease locus. To perform the analysis, you would use the following statements:

```
proc family data=example prefix=Marker outstat=stats tdt contcorr;
   id ped indiv father mother;
   trait disease / affected=2;
   var a1-a4;
run;

proc print data=stats;
   format ProbTDT pvalue6.5;
run;
```

This code creates an output data set stats, which contains the chi-square statistic, degrees of freedom, and p-value for testing each marker for linkage and association with the disease locus using the TDT. The PREFIX= option in the PROC FAMILY statement specifies that the two markers be named Marker1 and Marker2 in the output data set. The CONTCORR option indicates that the continuity correction of 0.5 should be used in calculating the chi-square statistic. The AFFECTED= option of the

TRAIT statement specifies which value of the variable disease should be considered "affected." Note that the pedigree ID variable is listed in the ID statement; however, it is not necessary for this data set, since all the individual IDs are unique. The same results would be obtained if this variable were omitted.

Here is the output data set that is produced:

Obs	Locus	ChiSq TDT	df TDT	Prob TDT
1	Marker1	1.57143	6	0.9546
2	Marker2	5.79861	5	0.3263

Figure 4.1. Statistics for the TDT

Figure 4.1 displays the statistics for the TDT. Since both markers are multiallelic, a joint test of all alleles at each marker is performed by default. The degrees of freedom (in the dfTDT column) indicate that there are seven alleles at Marker1 and six alleles at Marker2, since df$= k - 1$ where k is the number of marker alleles. The ProbTDT column shows that neither of the markers is significantly linked and associated with the disease locus.

Syntax

The following statements are available in PROC FAMILY.

PROC FAMILY < *options* > **;**
BY *variables* **;**
ID *variables* **;**
TRAIT *variable* </ **AFFECTED=***value*> **;**
VAR *variables* **;**

Items within angle brackets (< >) are optional, and statements following the PROC FAMILY statement can appear in any order. The ID, TRAIT, and VAR statements are required. The syntax of each statement is described in the following section in alphabetical order after the description of the PROC FAMILY statement.

PROC FAMILY Statement

PROC FAMILY < *options* > **;**

You can specify the following options in the PROC FAMILY statement.

COMBINE
specifies that the combined versions of the S-TDT and SDT be performed. Thus, families containing parental genotypes can be analyzed under certain conditions using the TDT, and otherwise the specified sibling test is performed. Note that if TDT is also being performed, the TDT is done independently of any other tests. By default, the combined versions are not used.

CONTCORR
CC
specifies that a continuity correction of 0.5 should be used for the TDT, S-TDT, and RC-TDT tests in their asymptotic normal approximations. By default, no correction is used.

DATA=*SAS-data-set*
names the input SAS data set to be used by PROC FAMILY. The default is to use the most recently created data set.

MULT= *JOINT* **|** *MAX*
specifies which multiallelic version of the TDT, S-TDT, SDT, and RC-TDT tests should be performed. The joint version of the multiallelic tests combines the analyses for each allele at a marker into one overall test statistic, with degrees of freedom (df) corresponding to the number of alleles at the marker. The max version of the multiallelic tests determines if there is at least one allele with a significant test statistic, using the maximum 1 df statistic over all alleles with a multiple testing adjustment made. By default, the joint version of the multiallelic tests is performed. This option has no effect on biallelic markers.

NDATA=*SAS-data-set*
names the input SAS data set containing names, or identifiers, for the markers used in the output. There must be a NAME variable in this data set, which should contain the same number of rows as there are markers in the input data set specified in the DATA= option. When there are fewer rows than there are markers, markers without a name are named using the PREFIX= option. Likewise, if there is no NDATA= data set specified, the PREFIX= option is used.

OUTSTAT=*SAS-data-set*
names the output SAS data set containing the p-values for the tests specified in the PROC FAMILY statement. When this option is omitted, an output data set is created by default and named according to the DATA*n* convention.

PREFIX=*prefix*
specifies a prefix to use in constructing names for marker variables in all output. For example, if PREFIX=VAR, the names of the variables are VAR1, VAR2, ..., VAR*n*. Note that this option is ignored when the NDATA= option is specified, unless there are fewer names in the NDATA data set than there are markers. If this option is omitted, PREFIX=M is the default.

RCTDT
requests that the reconstruction-combined TDT (RC-TDT) be performed. If none of the four test options (RCTDT, SDT, STDT, or TDT) are specified, then all four tests are performed by default.

SDT
requests that the SDT, a nonparametric alternative to the S-TDT, be performed. If none of the four test options (RCTDT, SDT, STDT, or TDT) are specified, then all four tests are performed by default. The COMBINE option can be used with this test to indicate that the combined version of the SDT should be performed.

STDT
requests that the sibling TDT (S-TDT), which analyzes data from sibships, be performed. If none of the four test options (RCTDT, SDT, STDT, or TDT) are specified, then all four tests are performed by default. The COMBINE option can be used with this test to indicate that the combined version of the S-TDT should be performed.

TDT
requests that the original TDT be performed. If none of the four test options (RCTDT, SDT, STDT, or TDT) are specified, then all four tests are performed by default.

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC FAMILY to obtain separate analyses on observations in groups defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables. The *variables* are one or more variables in the input data set.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data using the SORT procedure with a similar BY statement.
- Specify the BY statement option NOTSORTED or DESCENDING in the BY statement for the FAMILY procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables using the DATASETS procedure (in Base SAS software).

For more information on the BY statement, refer to the discussion in *SAS Language Reference: Concepts*. For more information on the DATASETS procedure, refer to the discussion in the *SAS Procedures Guide*.

ID Statement

ID *variables* **;**

The ID statement is required and must contain, in the following order, either:

- the pedigree ID, the individual ID, then the two parental ID variables, or
- the individual ID, then the two parental IDs

Thus if only three variables are specified in the ID statement, it is assumed that the pedigree identifier has been omitted. The pedigree ID is not necessary if all the individual identifiers are unique. The individual and two parental ID variables can be either numeric or character, but all three must be of the same type. The pedigree variable, if specified, can be either numeric or character regardless of the type of the other three identifiers.

TRAIT Statement

TRAIT *variable </ AFFECTED=value>* **;**

The TRAIT statement is required and identifies the trait variable. This variable must be binary, but may be either character or numeric. By default, the second value of the TRAIT variable that appears in the input data set is considered to be "affected" for the tests. If you would like to specify a different value for "affected," you may do so by adding the /AFFECTED=value option to the TRAIT statement. For a variable with a numeric format, the number that corresponds to "affected" should be specified (AFFECTED=1); if the variable has a character format, the level that corresponds to "affected" should be specified in quotes (AFFECTED="a").

VAR Statement

VAR *variables* **;**

The VAR statement identifies the variables containing the marker alleles. The VAR statement should contain $2m$ variable names, where m is the number of markers in the data set. Note that alleles for the same marker must be listed consecutively.

Details

Statistical Computations

For all tests, it is assumed that the marker has two alleles, M_1 and M_2. Extensions to multiallelic markers are made by performing the tests on each allele in turn, with the current allele being considered to be M_1 and all other alleles considered to be M_2. When the CONTCORR option is specified in the PROC FAMILY statement, the z score statistics of all versions of the TDT, S-TDT, and RC-TDT can be continuity corrected by subtracting 0.5 from the absolute value of the numerator. The two-sided p-value for each z score using the normal distribution is equivalent to using the p-value from the χ^2_1 distribution for the square of the z score, and this chi-square form of the statistic is reported in the output data set.

TDT

The TDT (Spielman, McGinnis, and Ewens 1993) is implemented using a normal approximation. This test includes families where both parents have been genotyped for the marker and at least one is heterozygous. If only one parent has been genotyped, that parent is heterozygous, and the affected child is not homozygous and does not have the same genotype as the typed parent, then the TDT can be applied to this family as well (Curtis and Sham 1995). The TDT tests for equality between the proportion of times a heterozygous parent transmits the M_1 allele to an affected child and the proportion of times a heterozygous parent transmits the M_2 allele to an affected child. The normal approximation to the binomial is used to form the z score statistic

$$Z = \frac{b - \frac{b+c}{2}}{\sqrt{\frac{b+c}{4}}}$$

where b is the number of M_1 alleles in all affected children from heterozygous parents and c is the number of M_2 alleles in affected children from heterozygous parents.

Two extensions to a multiallelic TDT are available. The first, which is performed by default or when MULT=JOINT is specified in the PROC FAMILY statement, combines the TDT for each of k alleles at a marker into one statistic as follows (Spielman and Ewens 1996):

$$T_J = \frac{k-1}{k} \sum_{v=1}^{k} Z_v^2$$

where Z_v is simply the Z defined in the preceding paragraph, with allele M_v treated as M_1 and all other alleles as M_2 for each $v = 1, ..., k$. T_J and the continuity-corrected form T'_J have an asymptotic χ^2_{k-1} distribution, and the corresponding p-value is reported.

Alternatively, if the MULT=MAX option is specified, either z_m or z'_m (when the CONTCORR option is specified) is used, where $z_m = \max_{1 \le v \le k} |Z_v|$. The equivalent one degree of freedom chi-square statistic is reported, and a Bonferroni correction is applied to its p-value.

Note: The TDT is a valid test of linkage and association only when the data consist of unrelated nuclear families and each family contains only one affected child. Otherwise, it is a valid test of linkage only.

S-TDT

The z score procedure given by Spielman and Ewens (1998) is used to calculate p-values for the S-TDT. This test can be applied to families where there are at least one affected sibling and one unaffected sibling, and not all siblings have the same genotype. The z score, whose two-sided p-value is approximated using the normal distribution, is calculated as $z = (Y - A)/\sqrt{V}$. Y represents the total observed number of M_1 alleles in the affected siblings. For t total siblings in the family, a affected and u unaffected, and r that are M_1/M_1 and s that are M_1/M_2, summing over families gives

$$A = \sum (2r + s)a/t$$

and

$$V = \sum au[4r(t - r - s) + s(t - s)]/[t^2(t - 1)]$$

as the expected value and variance of Y respectively.

When the COMBINE option is specified in the PROC FAMILY statement, the S-TDT and TDT are combined as follows: the TDT is applied to all families that meet the requirements described in the preceding section. The S-TDT is then applied to the remaining families that meet its requirements described in the preceding paragraph.

Using the notation already given for these tests, the z score for the combined test can then be written as

$$z = \frac{(Y + b) - (A + \frac{b+c}{2})}{\sqrt{V + \frac{b+c}{4}}}$$

For multiallelic markers, the same extensions can be made to the S-TDT and combined S-TDT that were made to the TDT (Monks, Kaplan, and Weir 1998); that is, either a joint test over all alleles, or the maximum z score of all the alleles with the p-value being Bonferroni-corrected.

Note: The S-TDT is a valid test of linkage and association only when the data consist of unrelated nuclear families and each family contains only one affected and one unaffected sibling. Otherwise, it is a valid test of linkage only.

SDT

The SDT (Horvath and Laird 1998) is a sign test used on discordant sibling pairs. As with the S-TDT, one affected sibling and one unaffected sibling are required to be in each family, but unlike the S-TDT, the SDT remains a valid test of linkage and association when the sibship is larger.

Continuing the notation from the previous tests, for a affected siblings in a family and u unaffected siblings in a family, treating each allele M_v in turn as M_1 and all other alleles as M_2, $v = 1, ..., k$, define for each family in the data the average number of v alleles among affected siblings and unaffected siblings respectively as

$$m_v^a = Y/a$$

$$m_v^u = [(2r + s) - Y]/u$$

Then $d_v = m_v^a - m_v^u$ for each family, and summing over families gives $S_v = \sum \text{sgn}(d_v)$, where $\text{sgn}(d_v) = 1$ for $d_v > 0$, 0 for $d_v = 0$, and -1 for $d_v < 0$. The joint multiallelic SDT statistic is then defined as $T = \mathbf{S'W^{-1}S}$ where $\mathbf{S'} = (S_1, ..., S_{k-1})'$ and $W_{vw} = \sum \text{sgn}(d_v)\text{sgn}(d_w)$, where $v, w = 1, ..., k-1$ (the information on the kth allele is not needed). This statistic has an asymptotic χ^2_{k-1} distribution, and this distribution is used to obtain p-values for the SDT. When there are only two alleles at the marker, this joint multiallelic version of the SDT reduces to the biallelic version of the SDT.

This sibship test is also combined with the TDT when the COMBINE option in the PROC FAMILY statement is specified, creating a test which can potentially use more of the data (Horvath and Laird 1998; Curtis, Miller, and Sham 1999). In order to maintain the test's validity as a test of association in families with more than one affected and one unaffected sibling, a nonparametric multiallelic TDT is used, which is in the same $\mathbf{S'W^{-1}S}$ form as the SDT. This test statistic for the joint test also has an asymptotic χ^2_{k-1} distribution, and the corresponding p-value is reported.

When the MULT=MAX option is specified in the PROC FAMILY statement, then the SDT chi-square statistic is simply $\max_{1 \le v \le k}(S_v^2 W_{vv}^{-1})$ and has one degree of freedom. This applies to the SDT when used alone or combined with the TDT. As with the other tests, a Bonferroni correction is made to the p-value.

RC-TDT

The RC-TDT (Knapp 1999) takes the combined S-TDT a step further by reconstructing missing parental genotypes when possible in order to use more families. The RC-TDT can be applied to families with at least one affected child that meet one of the following conditions:

- Both parents are typed with at least one heterozygous for M_1.
- One parent is typed, the other can be reconstructed, and at least one parent is heterozygous for M_1.
- Both parents' genotypes are missing but can be reconstructed, and at least one parent is heterozygous for M_1.
- At least one parental genotype is missing and cannot be reconstructed, but the conditions for the S-TDT are met.

As with the S-TDT, a z score is created using the statistic Y, but Knapp calculates a different expected value e and variance v of Y, which takes into account the bias created by the genotype reconstruction, to form the z score over all families:

$$z = (Y - e)/\sqrt{v}$$

For multiallelic markers, the same extensions can be made to the RC-TDT that were made to the TDT and S-TDT; that is, either a joint test over all alleles, or the maximum z score of all the alleles with the p-value being Bonferroni-corrected.

Note: The RC-TDT is a valid test of linkage and association only when the data consist of unrelated nuclear families and each family contains only one affected and one unaffected sibling. Otherwise, it is a valid test of linkage only.

Missing Values

An individual's genotype for a marker is considered missing if at least one of the alleles at the marker is missing. Any missing genotypes are excluded from all calculations. However, the individual's nonmissing genotypes at other loci can be used as part of the calculations. If a child has a missing trait value, then that individual is excluded from all calculations. However, missing trait values of individuals used only as parents do not affect the analysis. See the following section for information on missing values in the ID variables.

DATA= Data Set

The DATA= data set has columns representing markers, ID variables, and a trait, and rows representing the individuals. There must be one binary trait variable listed in the TRAIT statement; the three ID variables consisting of the individual's ID and the two parental IDs, all of the same type, must be listed in the ID statement, and optionally the pedigree ID if the individual identifiers are not unique. Note that only individuals with both parents appearing in the data, even if all the parents' genotypes are missing, can be used as affected children or in sib-pairs for analysis. However, if the individual is used only as a parent, then that individual's parents need not appear in the data. Also, if a pedigree ID variable is specified in the ID statement, any individual with a missing value for that variable is excluded from the analysis, as a parent and as a child. There are two columns for each marker, representing the two alleles at that marker carried by the individual. These two columns must be listed consecutively in the VAR statement. These marker variables must all be of the same type, but can be either character or numeric variables.

OUTSTAT= Data Set

The OUTSTAT= data set contains the following variables:

- the BY variables, if any
- Locus
- the chi-square statistics for each test performed: ChiSqTDT, ChiSqSTDT, ChiSqSDT, and ChiSqRCTDT
- the degrees of freedom for each test performed: dfTDT, dfSTDT, dfSDT, and dfRCTDT
- the p-values for each test performed: ProbTDT, ProbSTDT, ProbSDT, and ProbRCTDT

Example

Example 4.1. Performing Tests with Missing Parental Data

The following data are from GAW9 (Hodge 1995) and contain 20 nuclear families that are genotyped at two markers. The data have been modified so that each mother's genotype is missing.

```
data gaw;
   input ped id f_id m_id sex disease m11 m12 m21 m22;
   datalines;
 1    1    0    0 1 1  7  8  7  2
 1    2    0    0 2 1  .  .  .  .
 1  401    1    2 1 1  7  2  7  6
 1  402    1    2 1 1  8  2  7  6
 1  403    1    2 1 1  7  2  2  7
 1  404    1    2 2 2  8  2  7  7
```

```
 2    3    0    0 1 1  4  4  1  3
 2    4    0    0 2 1  .  .  .  .
 2  405    3    4 2 1  4  6  1  7
 2  406    3    4 2 2  4  4  3  7
 3    5    0    0 1 1  6  7  7  2
 3    6    0    0 2 1  .  .  .  .
 3  407    5    6 2 2  7  4  7  7
 4    7    0    0 1 1  1  8  7  3
 4    8    0    0 2 1  .  .  .  .
 4  408    7    8 2 2  8  4  7  3
 4  409    7    8 1 1  1  2  3  3
 4  410    7    8 2 1  8  2  7  3
 4  411    7    8 1 1  8  2  7  5
 5    9    0    0 1 1  7  1  6  2
 5   10    0    0 2 1  .  .  .  .
 5  412    9   10 2 2  7  6  6  2
 5  413    9   10 1 1  1  6  6  2
 6   11    0    0 1 1  8  4  2  3
 6   12    0    0 2 1  .  .  .  .
 6  414   11   12 1 2  8  1  2  7
 6  415   11   12 1 1  8  6  3  7
 7   13    0    0 1 1  4  6  2  2
 7   14    0    0 2 1  .  .  .  .
 7  416   13   14 1 1  4  5  2  7
 7  417   13   14 2 2  6  4  2  7
 7  418   13   14 2 1  6  5  2  6
 7  419   13   14 1 1  6  5  2  6
 8   15    0    0 1 1  6  8  2  7
 8   16    0    0 2 1  .  .  .  .
 8  420   15   16 2 1  6  2  7  7
 8  421   15   16 2 1  8  6  2  7
 8  422   15   16 2 2  6  6  7  7
 8  423   15   16 2 1  6  6  7  7
 9   17    0    0 1 2  4  7  2  7
 9   18    0    0 2 1  .  .  .  .
 9  424   17   18 2 2  4  5  7  2
 9  425   17   18 2 1  7  4  2  7
 9  426   17   18 1 1  4  5  2  2
10   19    0    0 1 1  6  4  2  7
10   20    0    0 2 1  .  .  .  .
10  427   19   20 2 2  4  4  7  2
11   21    0    0 1 1  4  7  7  7
11   22    0    0 2 1  .  .  .  .
11  428   21   22 1 1  7  6  7  2
11  429   21   22 2 2  7  4  7  2
11  430   21   22 2 1  7  6  7  3
12   23    0    0 1 1  7  6  7  5
12   24    0    0 2 1  .  .  .  .
12  431   23   24 1 2  6  4  7  7
13   25    0    0 1 1  4  1  2  8
13   26    0    0 2 1  .  .  .  .
13  432   25   26 1 1  4  8  2  6
13  433   25   26 1 2  1  8  8  6
13  434   25   26 1 1  1  4  2  6
```

```
14   27    0    0 1 1  7  6  3  2
14   28    0    0 2 1  .  .  .  .
14  435   27   28 1 1  6  2  3  3
14  436   27   28 1 1  7  4  3  7
14  437   27   28 1 1  6  2  2  7
14  438   27   28 1 1  7  4  2  7
14  439   27   28 2 2  6  2  2  7
14  440   27   28 1 1  6  4  3  7
15   29    0    0 1 1  2  4  7  4
15   30    0    0 2 1  .  .  .  .
15  441   29   30 1 1  4  2  7  7
15  442   29   30 2 2  4  8  4  7
15  443   29   30 2 1  4  2  7  5
15  444   29   30 2 1  4  2  7  5
15  445   29   30 1 1  2  8  7  5
;
```

Since there are missing parental data, the original TDT may not be the best test to perform on this data set. The following analysis uses the S-TDT, SDT, and RC-TDT to test markers for linkage with the disease locus.

```
proc family data=gaw prefix=Marker sdt stdt rctdt;
   id id f_id m_id;
   var m11 m12 m21 m22;
   trait disease / affected=2;
run;

proc print;
run;
```

The output data set, which is created by default, is displayed in Output 4.1.1.

Output 4.1.1. Output Data Set from PROC FAMILY

Obs	Locus	ChiSq STDT	ChiSq SDT	ChiSq RCTDT	df STDT	df SDT	df RCTDT	Prob STDT	Prob SDT	Prob RCTDT
1	Marker1	5.6179	3.9668	4.7398	6	6	6	0.467	0.681	0.578
2	Marker2	12.6191	10.6522	11.9388	7	7	7	0.082	0.155	0.103

Since only one parent is missing genotype information in each nuclear family, the TDT might be applicable to some of the families. The COMBINE option can be specified to use the TDT in the appropriate families, and the S-TDT or SDT for all other families. This option does not apply to the RC-TDT, so that test is omitted from this analysis.

```
proc family data=gaw prefix=Marker tdt sdt stdt combine;
   id id f_id m_id;
   var m11 m12 m21 m22;
   trait disease / affected=2;
```

```
run;

proc print;
run;
```

The output data set is displayed in Output 4.1.2.

Output 4.1.2. Output Data Set from PROC FAMILY Using COMBINE Option

Obs	Locus	ChiSq TDT	ChiSq STDT	ChiSq SDT	df TDT	df STDT	df SDT	Prob TDT	Prob STDT	Prob SDT
1	Marker1	8.00000	12.0511	5.7200	4	6	6	0.092	0.061	0.455
2	Marker2	2.66667	10.1802	10.6667	2	6	7	0.264	0.117	0.154

Note that the test statistics for the TDT and the S-TDT and SDT are not the same; this implies that not all families meet the requirements for the TDT. In this case, the S-TDT, SDT, and RC-TDT use more of the data than the TDT alone. However, since there is only one affected child in each nuclear family, the TDT is a valid test of association; since there is at least one occasion when there is more than one unaffected child in a nuclear family, the S-TDT and RC-TDT are not valid for testing for association of the marker with the disease locus (the SDT is always a valid test of association when the data consist of unrelated nuclear families). Both of these considerations, the amount of information that can be used and the validity for testing association, should be taken into account when deciding which test(s) to perform.

Another type of analysis can be performed using the MULT=MAX option in the PROC FAMILY statement. This option indicates that instead of doing a joint test over all the alleles at each marker, perform a test to see if any of the alleles at a marker are significantly linked with the disease locus. This analysis is invoked with the following code, using only the SDT and RC-TDT:

```
proc family data=gaw prefix=Marker sdt rctdt combine mult=max;
   id id f_id m_id;
   var m11 m12 m21 m22;
   trait disease / affected=2;
run;

proc print;
run;
```

The output data set produced by this code is displayed in Output 4.1.3.

Output 4.1.3. Output Data Set from PROC FAMILY Using MULT=MAX Option

Obs	Locus	ChiSq SDT	ChiSq RCTDT	df SDT	df RCTDT	Prob SDT	Prob RCTDT
1	Marker1	2.00000	2.90050	1	1	1.0000	0.6199
2	Marker2	2.66667	3.86422	1	1	0.8198	0.3946

The chi-square statistics for the tests always have one degree of freedom when the MULT=MAX option is used. Note, however, that the p-values are not the corresponding right-tailed probabilities for a χ^2_1 statistic; this is because the p-values are Bonferroni-corrected in order to account for taking the maximum of several chi-square statistics.

References

Curtis, D., Miller, M.B., and Sham, P.C. (1999), "Combining the Sibling Disequilibrium Test and Transmission/Disequilibrium Test for Multiallelic Markers," *American Journal of Human Genetics*, 64, 1785–1786.

Curtis, D. and Sham, P.C. (1995), "A Note on the Application of the Transmission Disequilibrium Test When a Parent is Missing," *American Journal of Human Genetics*, 56, 811–812.

Hodge, S.E. (1995), "An Oligogenic Disease Displaying Weak Marker Associations: A Summary of Contributions to Problem 1 of GAW9," *Genetic Epidemiology*, 12, 545–554.

Horvath, S. and Laird, N.M. (1998), "A Discordant-Sibship Test for Disequilibrium and Linkage: No Need for Parental Data," *American Journal of Human Genetics*, 63, 1886–1897.

Knapp, M. (1999), "The Transmission/Disequilibrium Test and Parental-Genotype Reconstruction: The Reconstruction-Combined Transmission/Disequilibrium Test," *American Journal of Human Genetics*, 64, 861–870.

Monks, S.A., Kaplan, N.L., and Weir, B.S. (1998), "A Comparative Study of Sibship Tests of Linkage and/or Association," *American Journal of Human Genetics*, 63, 1507–1516.

Spielman, R.S. and Ewens, W.J. (1996), "The TDT and Other Family-based Tests for Linkage Disequilibrium and Association," *American Journal of Human Genetics*, 59, 983–989.

Spielman, R.S. and Ewens, W.J. (1998), "A Sibship Test for Linkage in the Presence of Association: The Sib Transmission/Disequilibrium Test," *American Journal of Human Genetics*, 62, 450–458.

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)," *American Journal of Human Genetics*, 52, 506–516.

Chapter 5
The HAPLOTYPE Procedure

Chapter Contents

Chapter 5
The HAPLOTYPE Procedure

Overview

A *haplotype* is a combination of alleles at multiple loci on a single chromosome. A pair of haplotypes constitutes the multilocus genotype. Haplotype information has to be inferred as data are usually collected at the genotypic, not haplotype pair, level. For homozygous markers, there is no problem. If one locus has alleles A and a, and a second locus has alleles B and b, the observed genotype $AABB$ must contain two haplotypes of type AB; genotype $AaBB$ must contain haplotypes AB and aB, and so on. Haplotypes and their frequencies can be obtained directly. When both loci are heterozygous, however, there is ambiguity; a variety of combinations of haplotypes can generate the genotype, and it is not possible to determine directly which two haplotypes constitute any individual genotype. For example, the genotype $AaBb$ may be of type AB/ab with haplotypes AB and ab, or of type Ab/aB with haplotypes Ab and aB. The HAPLOTYPE procedure uses the expectation-maximization (EM) algorithm to generate maximum likelihood estimates of haplotype frequencies given a multilocus sample of genetic marker genotypes under the assumption of Hardy-Weinberg equilibrium (HWE). These estimates can then in turn be used to assign the probability that each individual possesses a particular haplotype pair.

Estimation of haplotype frequencies is important for several applications in genetic data analysis. One application is determining whether there is linkage disequilibrium (LD), or association, between loci. PROC HAPLOTYPE performs a likelihood ratio test to test the hypothesis of no LD between marker loci. Another application is association testing of disease susceptibility. Since sites that affect disease status are embedded in haplotypes, it has been postulated that the power of case-control studies might be increased by testing for haplotype rather than allele or genotype associations. One reason is that haplotypes might include two or more causative sites whose combined effect is measurable, particularly if they show synergistic interaction. Another is that fewer tests need be performed, although if there are a large number of haplotypes, this advantage is offset by the increased degrees of freedom of each test. PROC HAPLOTYPE can use case-control data to calculate test statistics for the hypothesis of no association between alleles comprising the haplotypes and disease status; such tests are carried out over all haplotypes at the loci specified, or for individual haplotypes.

Getting Started

Example

Assume you have a random sample with 25 individuals genotyped at four markers. You want to infer the gametic phases of the genotypes and estimate their frequencies. There are eight columns of data, with the first two columns containing the pair of alleles at the first marker, and the next two columns containing the pair of alleles for the second marker, and so on. Each row represents an individual. The data can be read into a SAS data set as follows:

```
data markers;
   input (m1-m8) ($);
   datalines;
B  B  A  B  B  B  A  A
A  A  B  B  A  B  A  B
B  B  A  A  B  B  B  B
A  B  A  B  A  B  A  B
A  A  A  B  A  B  B  B
B  B  A  A  A  B  A  B
A  B  B  B  A  B  A  A
A  B  A  A  A  A  A  A
B  B  A  A  A  A  A  B
A  B  A  B  A  B  B  B
A  B  A  B  A  B  A  A
B  B  A  B  A  B  A  A
A  B  A  A  A  B  A  B
A  B  B  B  B  B  A  B
A  A  A  B  A  A  A  B
B  B  A  B  A  B  A  B
A  B  B  B  A  A  A  B
B  B  B  B  A  A  A  A
A  B  A  A  A  B  A  A
A  B  A  A  A  B  A  B
B  B  A  A  A  A  A  B
A  A  A  B  A  A  A  B
A  B  A  A  A  A  B  B
A  A  A  A  A  A  A  A
A  B  B  B  A  A  A  A
;
```

You can now use PROC HAPLOTYPE to infer the possible haplotypes and estimate the four-locus haplotype frequencies in this sample. The following statements will perform these calculations:

```
proc haplotype data=markers out=hapout init=random prefix=SNP;
   var m1-m8;
run;

proc print data=hapout noobs round;
run;
```

This analysis uses the EM algorithm to estimate the haplotype frequencies from the sample. The standard errors and a confidence interval are estimated, by default, under a binomial assumption for each haplotype frequency estimate. A more precise estimate of the standard error can be obtained through the jackknife process by specifying the option SE=JACKKNIFE in the PROC HAPLOTYPE statement, but it takes considerably more computations (see the "Methods for Estimating Standard Error" section on page 73 for more information). The option INIT=RANDOM indicates that initial haplotype frequencies are randomly generated, using a random seed created by the system clock since the SEED= option is omitted. The default confidence level 0.95 is used since the ALPHA= option of the PROC HAPLOTYPE statement was omitted. Also by default, the convergence criterion of 0.00001 must be satisfied for one iteration, and the maximum number of iterations is set to 100. The PREFIX= option requests that the four markers, indicated by the eight allele variables in the VAR statement, be named SNP1-SNP4.

The results from the procedure are as follows.

The HAPLOTYPE Procedure

Analysis Information

Loci Used	**SNP1 SNP2 SNP3 SNP4**
Number of Individuals	**25**
Number of Starts	**1**
Convergence Criterion	**0.00001**
Iterations Checked for Conv.	**1**
Maximum Number of Iterations	**100**
Number of Iterations Used	**15**
Log Likelihood	**-95.94742**
Initialization Method	**Random**
Random Number Seed	**51220**
Standard Error Method	**Binomial**
Haplotype Frequency Cutoff	**0**

Figure 5.1. Analysis Information for the HAPLOTYPE Procedure

Figure 5.1 displays information on several of the settings used to perform the HAPLOTYPE procedure as well as information on the EM algorithm. Note that you can obtain from this table the random seed that was generated by the system clock if you need to replicate this analysis.

Haplotype Frequencies

Number	Haplotype	Freq	Standard Error	95% Confidence Limits	
1	A-A-A-A	0.14302	0.05001	0.04500	0.24105
2	A-A-A-B	0.07527	0.03769	0.00140	0.14914
3	A-A-B-A	0.00000	0.00000	0.00000	0.00000
4	A-A-B-B	0.00000	0.00010	0.00000	0.00020
5	A-B-A-A	0.09307	0.04151	0.01173	0.17442
6	A-B-A-B	0.05335	0.03210	0.00000	0.11627
7	A-B-B-A	0.00002	0.00061	0.00000	0.00122
8	A-B-B-B	0.07526	0.03769	0.00140	0.14913
9	B-A-A-A	0.08638	0.04013	0.00772	0.16504
10	B-A-A-B	0.08792	0.04046	0.00863	0.16722
11	B-A-B-A	0.07921	0.03858	0.00359	0.15482
12	B-A-B-B	0.10819	0.04437	0.02122	0.19517
13	B-B-A-A	0.10098	0.04304	0.01662	0.18534
14	B-B-A-B	0.00000	0.00001	0.00000	0.00002
15	B-B-B-A	0.09732	0.04234	0.01433	0.18030
16	B-B-B-B	0.00000	0.00001	0.00000	0.00002

Figure 5.2. Haplotype Frequencies from the HAPLOTYPE Procedure

Figure 5.2 displays the possible haplotypes in the sample and their estimated frequencies with standard errors and the lower and upper limits of the 95% confidence interval.

ID	m1	m2	m3	m4	m5	m6	m7	m8	HAPLOTYPE1	HAPLOTYPE2	PROB
1	B	B	A	B	B	B	A	A	B-A-B-A	B-B-B-A	1.00
2	A	A	B	B	A	B	A	B	A-B-A-A	A-B-B-B	1.00
2	A	A	B	B	A	B	A	B	A-B-A-B	A-B-B-A	0.00
3	B	B	A	A	B	B	B	B	B-A-B-B	B-A-B-B	1.00
4	A	B	A	B	A	B	A	B	A-A-A-B	B-B-B-A	0.26
4	A	B	A	B	A	B	A	B	A-B-A-A	B-A-B-B	0.36
4	A	B	A	B	A	B	A	B	A-B-A-B	B-A-B-A	0.15
4	A	B	A	B	A	B	A	B	A-B-B-A	B-A-A-B	0.00
4	A	B	A	B	A	B	A	B	A-B-B-B	B-A-A-A	0.23
5	A	A	A	B	A	B	B	B	A-A-A-B	A-B-B-B	1.00
6	B	B	A	A	A	B	A	B	B-A-A-A	B-A-B-B	0.57
6	B	B	A	A	A	B	A	B	B-A-A-B	B-A-B-A	0.43
7	A	B	B	B	A	B	A	A	A-B-A-A	B-B-B-A	1.00
7	A	B	B	B	A	B	A	A	A-B-B-A	B-B-A-A	0.00
8	A	B	A	A	A	A	A	A	A-A-A-A	B-A-A-A	1.00
9	B	B	A	A	A	A	A	B	B-A-A-A	B-A-A-B	1.00
10	A	B	A	B	A	B	B	B	A-B-A-B	B-A-B-B	0.47
10	A	B	A	B	A	B	B	B	A-B-B-B	B-A-A-B	0.53
11	A	B	A	B	A	B	A	A	A-A-A-A	B-B-B-A	0.65
11	A	B	A	B	A	B	A	A	A-B-A-A	B-A-B-A	0.35
11	A	B	A	B	A	B	A	A	A-B-B-A	B-A-A-A	0.00
12	B	B	A	B	A	B	A	A	B-A-A-A	B-B-B-A	0.51
12	B	B	A	B	A	B	A	A	B-A-B-A	B-B-A-A	0.49
13	A	B	A	A	A	B	A	B	A-A-A-A	B-A-B-B	0.72
13	A	B	A	A	A	B	A	B	A-A-A-B	B-A-B-A	0.28
14	A	B	B	B	B	B	A	B	A-B-B-B	B-B-B-A	1.00
15	A	A	A	B	A	A	A	B	A-A-A-A	A-B-A-B	0.52
15	A	A	A	B	A	A	A	B	A-A-A-B	A-B-A-A	0.48
16	B	B	A	B	A	B	A	B	B-A-A-B	B-B-B-A	0.44
16	B	B	A	B	A	B	A	B	B-A-B-B	B-B-A-A	0.56
17	A	B	B	B	A	A	A	B	A-B-A-B	B-B-A-A	1.00
18	B	B	B	B	A	A	A	A	B-B-A-A	B-B-A-A	1.00
19	A	B	A	A	A	B	A	A	A-A-A-A	B-A-B-A	1.00
20	A	B	A	A	A	B	A	B	A-A-A-A	B-A-B-B	0.72
20	A	B	A	A	A	B	A	B	A-A-A-B	B-A-B-A	0.28
21	B	B	A	A	A	A	A	B	B-A-A-A	B-A-A-B	1.00
22	A	A	A	B	A	A	A	B	A-A-A-A	A-B-A-B	0.52
22	A	A	A	B	A	A	A	B	A-A-A-B	A-B-A-A	0.48
23	A	B	A	A	A	A	B	B	A-A-A-B	B-A-A-B	1.00
24	A	A	A	A	A	A	A	A	A-A-A-A	A-A-A-A	1.00
25	A	B	B	B	A	A	A	A	A-B-A-A	B-B-A-A	1.00

Figure 5.3. Output Data Set from the HAPLOTYPE Procedure

Figure 5.3 displays each individual's genotype with each of the possible haplotype pairs that can comprise the genotype, and the probability the genotype can be resolved into each of the possible haplotype pairs.

Syntax

The following statements are available in PROC HAPLOTYPE.

PROC HAPLOTYPE < *options* > **;**
BY *variables* **;**
TRAIT *variable* **;**
VAR *variables* **;**

Items within angle brackets (< >) are optional, and statements following the PROC HAPLOTYPE statement can appear in any order. Only the VAR statement is required. The syntax for each statement is described in the following section in alphabetical order after the description of the PROC HAPLOTYPE statement.

PROC HAPLOTYPE Statement

PROC HAPLOTYPE < *options* > **;**

You can specify the following options in the PROC HAPLOTYPE statement.

ALPHA=*number*
specifies that a confidence level of $100(1-$*number*$)\%$ is to be used in forming the confidence intervals for estimates of haplotype frequencies. The value of *number* must be between 0 and 1, inclusive, and 0.05 is used as the default value if it is not specified.

CONV=*number*
specifies the convergence criterion for iterations of the EM algorithm, where $0 <$ *number* ≤ 1. The iteration process is stopped when the ratio of the change in the log likelihoods to the former log likelihood is less than or equal to *number* for the number of consecutive iterations specified in the NLAG= option (or 1 by default), or after the number of iterations specified in the MAXITER= option has been performed. The default value is 0.00001.

CUTOFF=*number*
specifies a lower bound on a haplotype's estimated frequency in order for that haplotype to be included in the "Haplotype Frequencies" table. The value of *number* must be between 0 and 1, inclusive. By default, all possible haplotypes from the sample are included in the table.

DATA=*SAS-data-set*
names the input SAS data set to be used by PROC HAPLOTYPE. The default is to use the most recently created data set.

INIT=*LINKEQ* **|** *RANDOM* **|** *UNIFORM*
indicates the method of initializing haplotype frequencies to be used in the EM algorithm. INIT=LINKEQ initializes haplotype frequencies assuming linkage equilibrium by calculating the product of the frequencies of the alleles that comprise the haplotype. INIT=RANDOM initializes haplotype frequencies with random values

from a Uniform(0,1) distribution, and INIT=UNIFORM assigns equal frequency to all haplotypes. By default, INIT=LINKEQ.

ITPRINT
requests that the "Iteration History" table be displayed. This option is ignored if the NOPRINT option is specified.

LD
requests that haplotype frequencies be calculated under the assumption of no LD, in addition to being calculated using the EM algorithm. When this option is specified, the "Test for Allelic Associations" table is displayed, which contains statistics for the likelihood ratio test for allelic associations. This option is ignored if the NOPRINT option is specified.

MAXITER=*number*
specifies the maximum number of iterations to be used in the EM algorithm. The number must be a non-negative integer. Iterations are carried out until convergence is reached according to the convergence criterion, or *number* iterations have been performed. The default is MAXITER=100.

NDATA=*SAS-data-set*
names the input SAS data set containing names, or identifiers, for the markers used in the output. There must be a NAME variable in this data set, which should contain the same number of rows as there are markers in the input data set specified in the DATA= option. When there are fewer rows than there are markers, markers without a name are named using the PREFIX= option. Likewise, if there is no NDATA= data set specified, the PREFIX= option is used.

NLAG=*number*
specifies the number of consecutive iterations that must meet the convergence criterion specified in the CONV= option (0.00001 by default) for the iteration process of the EM algorithm to stop. The number must be a positive integer. If this option is omitted, one iteration must satisfy the convergence criterion by default.

NOPRINT
suppresses the display of the "Analysis Information," "Iteration History," "Haplotype Frequencies," and "Test for Allelic Associations" tables. Either the OUT= option, the TRAIT statement, or both must be used with the NOPRINT option.

NSTART=*number*
specifies the number of different starts used for the EM algorithm. When this option is specified, PROC HAPLOTYPE starts the iterations with different random initial values *number*-2 times as well as once with uniform frequencies for all the haplotypes and once using haplotype frequencies assuming linkage equilibrium (independence). Results on the analysis using the initial values that produce the best log likelihood are then reported. The number must be a positive integer. If this option is omitted or NSTART=1, only one start with initial frequencies generated according to the INIT= option is used.

OUT=*SAS-data-set*
names the output SAS data set containing the probabilities of each genotype being resolved into all of the possible haplotype pairs.

OUTCUT=*number*
specifies a lower bound on a haplotype pair's estimated probability given the individual's genotype in order for that haplotype pair to be included in the OUT= data set. The value of *number* must be between 0 and 1, inclusive. By default, *number* $= 0.00001$. In order to be able to view all possible haplotype pairs for an individual's genotype, OUTCUT=0 can be specified.

PREFIX=*prefix*
specifies a prefix to use in constructing names for marker variables in all output. For example, if PREFIX=VAR, the names of the variables are VAR1, VAR2, ..., VAR*n*. Note that this option is ignored when the NDATA= option is specified, unless there are fewer names in the NDATA data set than there are markers. If this option is omitted, PREFIX=M is the default.

SE=*BINOMIAL* **|** *JACKKNIFE*
specifies the standard error estimation method. There are two methods available: the BINOMIAL option, which gives a standard error estimator from a binomial distribution and is the default method, and the JACKKNIFE option, which requests that the jackknife procedure be used to estimate the standard error.

SEED=*number*
specifies the initial seed for the random number generator used for creating the initial haplotype frequencies when INIT=RANDOM. The value for *number* must be a positive integer; the computer clock time is the default. For more details about seed values, refer to *SAS Language Reference: Concepts*.

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC HAPLOTYPE to obtain separate analyses on observations in groups defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables. The *variables* are one or more variables in the input data set.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data using the SORT procedure with a similar BY statement.
- Specify the BY statement option NOTSORTED or DESCENDING in the BY statement for the HAPLOTYPE procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables using the DATASETS procedure (in Base SAS software).

For more information on the BY statement, refer to the discussion in *SAS Language Reference: Concepts*. For more information on the DATASETS procedure, refer to the discussion in the *SAS Procedures Guide*.

TRAIT Statement

TRAIT *variable* < / *options* > **;**

The TRAIT statement identifies the binary variable that indicates which individuals are cases and which are controls, or represents a dichotomous trait. This variable can be character or numeric, but must have only two nonmissing levels. When this statement is used, the "Test for Marker-Trait Association" table is included in the output.

There are two options you can specify in the TRAIT statement:

PERMUTATION= *number*
PERM=*number*
specifies the number of permutations to be used to calculate the empirical p-value of the haplotype case-control tests. This number must be a positive integer. By default, no permutations are used and the p-value is calculated using the chi-square test statistic. Note that this option can greatly increase the computation time.

TESTALL
specifies that each individual haplotype should be tested for association with the TRAIT variable. When this option is included in the TRAIT statement, the "Tests for Haplotype-Trait Association" table is included in the output.

VAR Statement

VAR *variables* **;**

The VAR statement identifies the variables containing the marker alleles. The VAR statement should contain $2m$ variable names, where m is the number of markers in the data set and $m > 1$. Note that the two columns containing alleles for the same marker must be listed consecutively.

Details

Statistical Computations

The EM Algorithm

The EM algorithm (Excoffier and Slatkin 1995; Hawley and Kidd 1995; Long, Williams, and Urbanek 1995) iteratively furnishes the maximum likelihood estimates (MLEs) of m-locus haplotype frequencies, for any integer $m > 1$, when a direct solution for the MLE is not readily feasible. The EM algorithm assumes HWE; it has been argued (Fallin and Schork 2000) that positive increases in the Hardy-Weinberg disequilibrium coefficient (toward excess heterozygosity) may increase the error of the EM estimates, but negative increases (toward excess homozygosity) do not demonstrate a similar increase in the error. The iterations start with assigning initial values

to the haplotype frequencies. When the INIT=RANDOM option is included in the PROC HAPLOTYPE statement, uniformly distributed random values are assigned to all haplotype frequencies; when INIT=UNIFORM, each haplotype is given an initial frequency of $1/h$, where h is the number of possible haplotypes in the sample. Otherwise, the product of the frequencies of the alleles that comprise the haplotype is used as the initial frequency for the haplotype. Different starting values can lead to different solutions since a maximum that is found could be a local maximum and not the global maximum. You can try different starting values for the EM algorithm by specifying a number greater than 1 in the NSTART= option to get better estimates. The expectation and maximization steps (E-step and M-step, respectively) are then carried out until the convergence criterion is met or the number of iterations exceeds the number specified in the MAXITER= option of the PROC HAPLOTYPE statement.

For a sample of n individuals, suppose the ith individual has genotype G_i. The probability of this genotype in the population is P_i, so the log likelihood is

$$\log L = \sum_{i=1}^{n} \log P_i$$

which is calculated after each iteration's E-step of the EM algorithm, described in the following paragraphs.

Let h_j be the jth possible haplotype and f_j its frequency in the population. For genotype G_i, the set H_i is the collection of pairs of haplotypes, h_j and its "complement" h_j^{ci}, that constitute that genotype. The haplotype frequencies f_j used in the E-step for iteration 0 of the EM algorithm are given by the initial values; all subsequent iterations use the haplotype frequencies calculated by the M-step of the previous iteration. The E-step sets the genotype frequencies to be products of these frequencies:

$$P_i = \sum_{j \in H_i} f_j f_j^{ci}$$

When G_i has m heterozygous loci, there are 2^{m-1} terms in this sum. The number of times haplotype h_j occurs in the sum is written as m_{ij}, which is 2 if G_i is completely homozygous, and either 1 or 0 otherwise.

The M-step sets new haplotype frequencies from the genotype frequencies:

$$f_j = \frac{1}{2n} \sum_{i=1}^{n} \frac{m_{ij} f_j f_j^{ci}}{P_i}$$

The EM algorithm increases the likelihood after each iteration, and multiple starting points can generally lead to the global maximum.

Once the EM algorithm has arrived at the MLEs of the haplotype frequencies, each individual i's probability of having a particular haplotype pair (h_j, h_j^{ci}) given the individual's genotype G_i is calculated as

$$\Pr\{h_j, h_j^{ci} | G_i\} = \frac{f_j f_j^{ci}}{P_i}$$

for each $j \in H_i$. These probabilities are displayed in the OUT= data set.

Methods for Estimating Standard Error

Typically, an estimate of the variance of a haplotype frequency is obtained by inverting the estimated information matrix from the distribution of genotype frequencies. However, it often turns out that in a large multilocus system, a certain proportion of haplotypes have ML frequencies equal or close to zero which makes the sample information matrix nearly singular (Excoffier and Slatkin 1995). Therefore, two approximation methods are used to estimate the variances, as proposed by Hawley and Kidd (1995).

The binomial method estimates the standard error by calculating the square root of the binomial variance, as if the haplotype frequencies are obtained by direct counting:

$$\text{Var}_\text{B}(f_j) = \frac{f_j(1 - f_j)}{2n - 1}$$

The jackknife method is a simulation-based method that can be used to estimate the standard errors of haplotype frequencies. Each individual is in turn removed from the sample, and all the haplotype frequencies are recalculated from this "delete-1" sample. Let $T_{n-1,i}$ be the haplotype frequency estimator from the ith "delete-1" sample; then the jackknife variance estimator has the following formula:

$$\text{Var}_\text{J}(f_j) = \frac{n-1}{n} \sum_{i=1}^{n} \left(T_{n-1,i} - \frac{1}{n} \sum_{j=1}^{n} T_{n-1,j} \right)^2$$

and the square root of this variance estimate is the estimate of standard error. The jackknife is less dependent on the model assumptions; however, it requires computing the statistic n times.

Confidence intervals with confidence level $1 - \alpha$ for the haplotype frequency estimates from the final iteration are then calculated using the following formula:

$$f_j \pm z_{1-\alpha/2} \sqrt{\text{Var}(f_j)}$$

where $z_{1-\alpha/2}$ is the value from the standard normal distribution that has a right-tail probability of $\alpha/2$.

Testing for Allelic Associations

When the LD option is specified in the PROC HAPLOTYPE statement, haplotype frequencies are calculated using the EM algorithm as well as by assuming no allelic associations among loci, that is, no LD. Under the null hypothesis of no LD, haplotype frequencies are simply the product of the individual allele frequencies. The log likelihood under the null hypothesis, $\log L_0$, is calculated based on these haplotype frequencies with degrees of freedom $\text{df}_0 = \sum_{i=1}^{m}(k_i - 1)$, where m is the number of loci and k_i is the number of alleles for the ith locus (Zhao, Curtis, and Sham 2000). Under the alternative hypothesis, the log likelihood, $\log L_1$ is calculated from the EM estimates of the haplotype frequencies with degrees of freedom $\text{df}_1 =$ number of haplotypes $- 1$. A likelihood ratio test is used to test this hypothesis as follows:

$$2(\log L_1 - \log L_0) \sim \chi^2_\nu$$

where $\nu = \text{df}_1 - \text{df}_0$ is the difference between the number of degrees of freedom under the null hypothesis and the alternative.

Testing for Trait Associations

When the TRAIT statement is included in PROC HAPLOTYPE, case-control tests are performed to test for association between the dichotomous trait (often, an indicator of individuals with or without a disease) and the marker loci using haplotypes. In addition to an omnibus test that is performed over all haplotypes, when the TESTALL option is specified in the TRAIT statement, a test for association between each individual haplotype and the trait is performed. Note that the individual haplotype tests should only be performed if the omnibus test statistic is significant.

Chi-Square Tests

The test performed over all haplotypes is based on the log likelihoods: under the null hypothesis, the log likelihood over all the individuals in the sample, regardless of the value of their trait variable, is calculated as described in the section "The EM Algorithm" on page 71; the log likelihood is also calculated separately for the two sets of individuals within the sample as determined by the trait value under the alternative hypothesis of marker-trait association. A likelihood ratio test (LRT) statistic can then be formed as follows:

$$X^2 = 2(\log L_1 + \log L_2 - \log L_0)$$

where $\log L_0$, $\log L_1$, and $\log L_2$ are the log likelihoods under the null hypothesis, for individuals with the first trait value, and for individuals with the second trait value, respectively (Zhao, Curtis, and Sham 2000). Defining degrees of freedom for each log likelihood similarly, this statistic has an asymptotic chi-square distribution with $(\text{df}_1 + \text{df}_2 - \text{df}_0)$ degrees of freedom.

An association between individual haplotypes and the trait can also be tested. To do so, the following contingency table is formed:

Table 5.1. Haplotype-Trait Counts

	Hap 1	Hap 2	Total
Trait 1	c_{11}	c_{12}	t_1
Trait 2	c_{21}	c_{22}	t_2
Total	h_1	h_2	T

where $T = 2n = t_1 + t_2 = h_1 + h_2$, the total number of haplotypes in the sample, "Hap 1" refers to the current haplotype being tested, "Hap 2" refers to all other haplotypes, and c_{ij} is the count of individuals with trait i and haplotype j (note that these counts are not necessarily integers since haplotypes are not observed; they are calculated based on the estimated haplotype frequencies). The usual contingency table chi-square test statistic

$$\sum_{i=1,2} \sum_{j=1,2} \frac{(c_{ij} - t_i h_j / T)^2}{t_i h_j / T}$$

has a 1 df chi-square distribution.

Exact Tests

Since the assumption of a chi-square distribution in the preceding section may not hold, exact p-values can also be calculated. New samples are formed by randomly permuting the trait values, and either of the chi-square test statistics shown in the previous section can be calculated for each of these samples. The number of new samples created is determined by the number given in the PERMUTATION= option of the TRAIT statement. The exact p-value is then calculated as m/p, where m is the number of samples with a test statistic greater than or equal to the test statistic in the actual sample and p is the total number of permutation samples. This method is used to obtain empirical p-values for both the overall and the individual haplotype tests (Zhao, Curtis, and Sham 2000; Fallin et al. 2001).

Missing Values

An individual's m-locus genotype is considered to be partially missing if any, but not all, of the alleles are missing. Genotypes with all missing alleles are dropped. Also, if there are any markers with all missing values in a BY group (or the entire data set if there is no BY statement), no calculations are performed for that BY group. Partially missing genotypes are used in the EM algorithm and the jackknife procedure. In calculating the allele frequencies, missing alleles are dropped and the frequency of an allele u at a marker is obtained as the number of u alleles in the data divided by the total number of non-missing alleles at the marker in the data. In the E-step of the EM algorithm, the frequency of a partially missing genotype is updated for every possible genotype. In the M-step, haplotypes resulting from a missing genotype may bear some missing alleles. Such a haplotype is not considered as a new haplotype, but rather all existing haplotypes that have alleles identical to the nonmissing alleles of this haplotype are updated. Dealing with missing genotypes involves looping through all possible genotypes in the E-step and all possible haplotypes in the M-step.

Depending on the input data set, missing genotypes can increase the computation time substantially.

When the TRAIT statement is specified, any observation with a missing trait value is dropped from calculations used in the tests for marker-trait association and haplotype-trait associations. However, observations with missing trait values are included in calculating the frequencies shown in the "Haplotype Frequencies" table, which are then used in the OUT= data set. The combined frequencies listed in the "Tests for Haplotype-Trait Association" table may therefore be different than these frequencies in this situation.

OUT= Data Set

The OUT= data set contains the following variables: the BY variables (if any), ID that identifies the individual, the $2m$ variables listed in the VAR statement, HAPLOTYPE1 and HAPLOTYPE2 that contain the pair of haplotypes of which each genotype can be comprised, and PROB containing the probability of each individual's genotype being resolved into that haplotype pair.

Displayed Output

This section describes the displayed output from PROC HAPLOTYPE. See the "ODS Table Names" section on page 78 for details about how this output interfaces with the Output Delivery System.

Analysis Information

The "Analysis Information" table lists information on the following settings used in PROC HAPLOTYPE:

- Loci Used, the loci used to form haplotypes
- Number of Individuals
- Number of Starts, the value specified in the NSTART= option or the default (1)
- Convergence Criterion, the value specified in the CONV= option or the default (0.00001)
- Iterations Checked for Conv., the value specified in the NLAG= option or the default (1)
- Maximum Number of Iterations, the value specified in the MAXITER= option or the default (100)
- Number of Iterations Used, as determined by the CONV= or MAXITER= option
- Log Likelihood, from the last iteration performed
- Initialization Method, the method specified in the INIT= option or "Linkage Equilibrium" by default
- Random Number Seed, the value specified in the SEED= option or generated by the system clock, or "Not Used" if INIT=LINKEQ or INIT=UNIFORM

- Standard Error Method, the method specified in the SE= option or "Binomial" by default
- Haplotype Frequency Cutoff, the value specified in the CUTOFF= option or the default (0)

Iteration History

The "Iteration History" table displays the log likelihood and the ratio of change for each iteration of the EM algorithm.

Haplotype Frequencies

The "Haplotype Frequencies" table lists all the possible m-locus haplotypes in the sample (where $2m$ variables are specified in the VAR statement), with an estimate of the haplotype frequency, the standard error of the frequency, and the lower and upper limits of the confidence interval for the frequency based on the confidence level determined by the ALPHA= option of the PROC HAPLOTYPE statement (0.95 by default). When the LD option is specified in the PROC HAPLOTYPE statement, haplotype frequency estimates are calculated both under the null hypothesis of no allelic association by taking the product of allele frequencies, and under the alternative, which allows for associations, using the EM algorithm.

Test for Allelic Associations

The "Test for Allelic Associations" table displays the degrees of freedom and log likelihood for the null hypothesis of no association and the alternative hypothesis of associations between markers. The chi-square statistic and its p-value are also shown for the test of these hypotheses.

Test for Marker-Trait Association

The "Test for Marker-Trait Association" table displays the number of observations, degrees of freedom, and log likelihood for both trait values as well as the combined sample. The chi-square test statistic and its corresponding p-value from performing the case-control test, testing the hypothesis of no association between the trait and the marker loci used in PROC HAPLOTYPE, are also given. When the PERMUTATION= option is included in the TRAIT statement, exact p-values are provided as well.

Tests for Haplotype-Trait Association

The "Tests for Haplotype-Trait Association" table displays statistics from case-control tests performed on each individual haplotype when the TESTALL option is included in the TRAIT statement. A significant p-value indicates that there is an association between the haplotype and the trait. When the PERMUTATION= option is also given in the TRAIT statement, exact p-values are provided as well.

ODS Table Names

PROC HAPLOTYPE assigns a name to each table it creates, and you must use this name to reference the table when using the Output Delivery System (ODS). These names are listed in the following table.

Table 5.2. ODS Tables Created by the HAPLOTYPE Procedure

ODS Table Name	Description	Statement or Option
AnalysisInfo	Analysis information	default
IterationHistory	Iteration history	ITPRINT
ConvergenceStatus	Convergence status	default
HaplotypeFreq	Haplotype frequencies	default
LDTest	Test for allelic associations	LD
CCTest	Test for marker-trait association	TRAIT statement
HapTraitTest	Tests for haplotype-trait association	TRAIT / TESTALL

Examples

Example 5.1. Estimating Three-Locus Haplotype Frequencies

Here is an example of 227 individuals genotyped at three markers, data which were created based on genotype frequency tables from the Lab of Statistical Genetics at Rockefeller University (2001). Note that when reading in the data, there are four individuals' genotypes per line, except for the last line of the DATA step, which contains three individuals' genotypes. The SAS data set that is created contains one individual per row with six columns representing the two alleles at each of three marker loci.

```
data ehdata;
   input m1-m6 @@;
   datalines;
1 1 1 1 1 3 1 1 1 1 1 3 1 1 1 1 1 3 1 1 1 1 1 3
1 1 1 1 1 3 1 1 1 1 1 3 1 1 1 1 1 3 1 1 1 1 1 3
1 1 1 1 1 3 1 1 1 1 1 3 1 1 1 1 1 3 1 1 1 1 1 3
1 1 1 1 2 3 1 1 1 1 2 3 1 1 1 1 2 3 1 1 1 1 3 3
1 1 1 1 3 3 1 1 1 1 3 3 1 1 1 1 3 3 1 1 1 1 3 3
1 1 1 1 3 3 1 1 1 1 3 3 1 1 1 1 3 3 1 1 1 1 3 3
1 1 1 2 1 1 1 1 1 2 1 1 1 1 1 2 1 2 1 1 1 2 1 2
1 1 1 2 1 2 1 1 1 2 1 2 1 1 1 2 1 2 1 1 1 2 1 2
1 1 1 2 1 2 1 1 1 2 1 2 1 1 1 2 1 2 1 1 1 2 1 2
1 1 1 2 1 2 1 1 1 2 1 2 1 1 1 2 2 2 1 1 1 2 2 2
1 1 1 2 1 3 1 1 1 2 1 3 1 1 1 2 2 3 1 1 1 2 2 3
1 1 1 2 2 3 1 1 1 2 3 3 1 1 1 2 3 3 1 1 1 2 3 3
1 1 1 2 3 3 1 1 2 2 1 1 1 1 2 2 1 1 1 1 2 2 1 1
1 1 2 2 1 1 1 1 2 2 1 1 1 1 2 2 1 2 1 1 2 2 1 2
1 1 2 2 1 2 1 1 2 2 1 2 1 1 2 2 2 2 1 1 2 2 2 2
```

```
1 1 2 2 2 2 1 1 2 2 2 2 1 1 2 2 1 3 1 1 2 2 1 3
1 1 2 2 3 3 1 1 2 2 3 3 1 1 2 2 3 3 1 1 2 2 3 3
1 1 2 2 3 3 1 1 2 2 3 3 1 1 2 2 3 3 1 1 2 2 3 3
1 1 2 2 3 3 1 1 2 2 3 3 1 1 2 2 3 3 1 1 2 2 3 3
1 1 2 2 3 3 1 1 2 2 3 3 1 1 2 2 3 3 1 1 2 2 3 3
1 1 2 2 3 3 1 1 2 2 3 3 1 2 1 1 1 1 1 2 1 1 1 1
1 2 1 1 1 1 1 2 1 1 1 1 1 2 1 1 1 1 1 2 1 1 1 1
1 2 1 1 1 1 1 2 1 1 1 2 1 2 1 1 1 2 1 2 1 1 1 3
1 2 1 1 1 3 1 2 1 1 2 3 1 2 1 1 2 3 1 2 1 1 2 3
1 2 1 1 2 3 1 2 1 1 2 3 1 2 1 1 2 3 1 2 1 1 3 3
1 2 1 1 3 3 1 2 1 1 3 3 1 2 1 2 1 1 1 2 1 2 1 1
1 2 1 2 1 1 1 2 1 2 1 1 1 2 1 2 1 1 1 2 1 2 1 1
1 2 1 2 1 1 1 2 1 2 1 1 1 2 1 2 1 1 1 2 1 2 1 3
1 2 1 2 1 3 1 2 1 2 1 3 1 2 1 2 2 3 1 2 1 2 2 3
1 2 1 2 2 3 1 2 1 2 2 3 1 2 1 2 3 3 1 2 1 2 3 3
1 2 2 2 1 1 1 2 2 2 1 1 1 2 2 2 1 1 1 2 2 2 1 1
1 2 2 2 1 1 1 2 2 2 1 1 1 2 2 2 1 1 1 2 2 2 1 1
1 2 2 2 1 1 1 2 2 2 1 1 1 2 2 2 1 2 1 2 2 2 1 2
1 2 2 2 1 2 1 2 2 2 1 3 1 2 2 2 1 3 1 2 2 2 2 3
1 2 2 2 2 3 1 2 2 2 2 3 1 2 2 2 3 3 1 2 2 2 3 3
1 2 2 2 3 3 1 2 2 2 3 3 1 2 2 2 3 3 1 2 2 2 3 3
1 2 2 2 3 3 1 2 2 2 3 3 2 2 1 1 1 1 2 2 1 1 1 2
2 2 1 1 1 2 2 2 1 1 2 2 2 2 1 1 2 2 2 2 1 1 2 2
2 2 1 1 2 2 2 2 1 1 2 2 2 2 1 1 2 2 2 2 1 1 2 2
2 2 1 1 2 2 2 2 1 1 2 2 2 2 1 1 2 2 2 2 1 1 1 3
2 2 1 1 1 3 2 2 1 1 2 3 2 2 1 1 2 3 2 2 1 1 2 3
2 2 1 1 2 3 2 2 1 1 2 3 2 2 1 1 2 3 2 2 1 1 2 3
2 2 1 1 2 3 2 2 1 1 2 3 2 2 1 1 3 3 2 2 1 1 3 3
2 2 1 1 3 3 2 2 1 1 3 3 2 2 1 2 1 1 2 2 1 2 1 1
2 2 1 2 1 1 2 2 1 2 1 1 2 2 1 2 1 1 2 2 1 2 1 2
2 2 1 2 1 2 2 2 1 2 1 2 2 2 1 2 2 2 2 2 1 2 2 2
2 2 1 2 2 2 2 2 1 2 2 2 2 2 1 2 1 3 2 2 1 2 1 3
2 2 1 2 1 3 2 2 1 2 1 3 2 2 1 2 1 3 2 2 1 2 1 3
2 2 1 2 2 3 2 2 1 2 2 3 2 2 1 2 2 3 2 2 1 2 2 3
2 2 1 2 2 3 2 2 1 2 2 3 2 2 1 2 2 3 2 2 1 2 3 3
2 2 1 2 3 3 2 2 1 2 3 3 2 2 2 2 1 1 2 2 2 2 1 1
2 2 2 2 1 1 2 2 2 2 1 1 2 2 2 2 1 1 2 2 2 2 1 1
2 2 2 2 1 1 2 2 2 2 1 1 2 2 2 2 1 1 2 2 2 2 1 2
2 2 2 2 1 2 2 2 2 2 1 2 2 2 2 2 2 3 2 2 2 2 2 3
2 2 2 2 2 3 2 2 2 2 3 3 2 2 2 2 3 3 2 2 2 2 3 3
2 2 2 2 3 3 2 2 2 2 3 3 2 2 2 2 3 3 2 2 2 2 3 3
2 2 2 2 3 3 2 2 2 2 3 3 2 2 2 2 3 3
;
```

The haplotype frequencies can be estimated using the EM algorithm and their standard errors estimated using the jackknife method by implementing the following code:

```
proc haplotype data=ehdata se=jackknife maxiter=20 itprint nlag=4;
   var m1-m6;
run;
```

This produces the following ODS output:

Output 5.1.1. Analysis Information for the HAPLOTYPE Procedure

The HAPLOTYPE Procedure

Analysis Information

Loci Used	M1 M2 M3
Number of Individuals	227
Number of Starts	1
Convergence Criterion	0.00001
Iterations Checked for Conv.	4
Maximum Number of Iterations	20
Number of Iterations Used	11
Log Likelihood	-934.97918
Initialization Method	Linkage Equilibrium
Standard Error Method	Jackknife
Haplotype Frequency Cutoff	0

Output 5.1.1 displays information on several of the settings used to perform the HAPLOTYPE procedure on the ehdata data set. Note that though the MAXITER= option was set to 20 iterations, convergence according to the criterion of 0.00001 was reached for four consecutive iterations prior to the 20th iteration, at which point the estimation process stopped. To obtain more precise frequency estimates, a lower convergence criterion can be used.

Output 5.1.2. Iteration History for the HAPLOTYPE Procedure

Iteration History

Iter	LogLike	Ratio Changed
0	-953.89697	
1	-937.92181	0.01675
2	-935.91870	0.00214
3	-935.35775	0.00060
4	-935.13050	0.00024
5	-935.03710	0.00010
6	-935.00051	0.00004
7	-934.98679	0.00001
8	-934.98180	0.00001
9	-934.98002	0.00000
10	-934.97940	0.00000
11	-934.97918	0.00000

Output 5.1.3. Convergence Status for the HAPLOTYPE Procedure

Algorithm converged.

Because the ITPRINT option was specified in the PROC HAPLOTYPE statement, the iteration history of the EM algorithm is included in the ODS output. Output 5.1.2 contains the table displaying this information. By default, the “Convergence Status” table is displayed (Output 5.1.3), which only consists of one line indicating whether

convergence was met.

Output 5.1.4. Haplotype Frequencies from the HAPLOTYPE Procedure

Haplotype Frequencies

Number	Haplotype	Freq	Standard Error	95% Confidence Limits	
1	1-1-1	0.09170	0.01505	0.06221	0.12119
2	1-1-2	0.02080	0.00952	0.00214	0.03946
3	1-1-3	0.11509	0.01766	0.08048	0.14971
4	1-2-1	0.07904	0.01696	0.04580	0.11228
5	1-2-2	0.06768	0.01546	0.03738	0.09799
6	1-2-3	0.12788	0.02094	0.08685	0.16891
7	2-1-1	0.05521	0.01227	0.03115	0.07926
8	2-1-2	0.11700	0.01782	0.08207	0.15193
9	2-1-3	0.07376	0.01495	0.04446	0.10307
10	2-2-1	0.11766	0.01831	0.08177	0.15355
11	2-2-2	0.03020	0.00899	0.01257	0.04782
12	2-2-3	0.10397	0.01833	0.06805	0.13989

Output 5.1.4 displays the 12 possible three-locus haplotypes in the data and their estimated haplotype frequencies, standard errors, and bounds for the 95% confidence intervals for the estimates.

To see how the CUTOFF= option affects the "Haplotype Frequencies" table, suppose you want to view only the haplotypes with an estimated frequency of at least 0.10. The following code creates such a table:

```
proc haplotype data=ehdata se=jackknife cutoff=0.10 nlag=4;
  var m1-m6;
run;
```

Now, the "Haplotype Frequencies" table is displayed as:

Output 5.1.5. Haplotype Frequencies from the HAPLOTYPE Procedure Using the CUTOFF= Option

The HAPLOTYPE Procedure

Haplotype Frequencies

Number	Haplotype	Freq	Standard Error	95% Confidence Limits	
1	1-1-3	0.11509	0.01766	0.08048	0.14971
2	1-2-3	0.12788	0.02094	0.08685	0.16891
3	2-1-2	0.11700	0.01782	0.08207	0.15193
4	2-2-1	0.11766	0.01831	0.08177	0.15355
5	2-2-3	0.10397	0.01833	0.06805	0.13989

Output 5.1.5 displays only the four three-locus haplotypes with estimated frequencies of at least 0.10. This option is especially useful for keeping the "Haplotype Frequencies" table to a manageable size when many marker loci or loci with several

alleles are used, and many of the haplotypes have estimated frequencies very near zero. Using CUTOFF=1 suppresses the “Haplotype Frequencies” table.

Example 5.2. Using Multiple Runs of the EM Algorithm

Continuing the example from the section “Getting Started” on page 64, suppose you are concerned that the likelihood reached a local and not a global maximum. You can request that PROC HAPLOTYPE use several different sets of initial haplotype frequencies to ensure that you find a global maximum of the likelihood. The following code invokes the EM algorithm using five different sets of initial values, including the set used in the Getting Started example:

```
proc haplotype data=markers prefix=SNP init=random seed=51220
               nstart=5;
   var m1-m8;
run;
```

The NSTART=5 option requests that the EM algorithm be run three times using randomly generated initial frequencies, including once using the seed 51220 that was previously used, once using uniform initial frequencies, and once using haplotype frequencies given by the product of the allele frequencies. The following two tables are from the run that produced the best log likelihood:

Output 5.2.1. Output from PROC HAPLOTYPE

The HAPLOTYPE Procedure

Analysis Information

Loci Used	SNP1 SNP2 SNP3 SNP4
Number of Individuals	25
Number of Starts	5
Convergence Criterion	0.00001
Iterations Checked for Conv.	1
Maximum Number of Iterations	100
Number of Iterations Used	15
Log Likelihood	-95.94738
Initialization Method	Random
Random Number Seed	23152
Standard Error Method	Binomial
Haplotype Frequency Cutoff	0

Haplotype Frequencies

Number	Haplotype	Freq	Standard Error	95% Confidence Limits	
1	A-A-A-A	0.14302	0.05001	0.04500	0.24105
2	A-A-A-B	0.07527	0.03769	0.00140	0.14914
3	A-A-B-A	0.00000	0.00000	0.00000	0.00000
4	A-A-B-B	0.00000	0.00010	0.00000	0.00020
5	A-B-A-A	0.09307	0.04151	0.01173	0.17442
6	A-B-A-B	0.05335	0.03210	0.00000	0.11627
7	A-B-B-A	0.00002	0.00061	0.00000	0.00122
8	A-B-B-B	0.07526	0.03769	0.00140	0.14913
9	B-A-A-A	0.08638	0.04013	0.00772	0.16504
10	B-A-A-B	0.08792	0.04046	0.00863	0.16722
11	B-A-B-A	0.07921	0.03858	0.00359	0.15482
12	B-A-B-B	0.10819	0.04437	0.02122	0.19517
13	B-B-A-A	0.10098	0.04304	0.01662	0.18534
14	B-B-A-B	0.00000	0.00001	0.00000	0.00002
15	B-B-B-A	0.09732	0.04234	0.01433	0.18030
16	B-B-B-B	0.00000	0.00001	0.00000	0.00002

Example 5.3. Testing for Linkage Disequilibrium

Again looking at the data from the Lab of Statistical Genetics at Rockefeller University (2001), if you request the test for linkage disequilibrium by specifying the LD option in the PROC HAPLOTYPE statement, the “Test for Allelic Associations” table containing the test statistics is included in the output.

```
proc haplotype data=ehdata ld;
   var m1-m6;
run;
```

The “Haplotype Frequencies” table now contains an extra column of the haplotype frequencies under the null hypothesis.

Output 5.3.1. Haplotype Frequencies Under the Null and Alternative Hypotheses

The HAPLOTYPE Procedure

Haplotype Frequencies

Number	Haplotype	H0 Freq	H1 Freq	Standard Error	95% Confidence Limits	
1	1-1-1	0.08172	0.09124	0.01353	0.06472	0.11775
2	1-1-2	0.05605	0.02124	0.00677	0.00796	0.03452
3	1-1-3	0.10006	0.11501	0.01499	0.08563	0.14439
4	1-2-1	0.09084	0.07952	0.01271	0.05461	0.10443
5	1-2-2	0.06231	0.06726	0.01177	0.04419	0.09032
6	1-2-3	0.11122	0.12794	0.01569	0.09718	0.15870
7	2-1-1	0.08100	0.05540	0.01075	0.03433	0.07647
8	2-1-2	0.05556	0.11690	0.01510	0.08732	0.14649
9	2-1-3	0.09918	0.07378	0.01228	0.04971	0.09785
10	2-2-1	0.09005	0.11746	0.01513	0.08781	0.14711
11	2-2-2	0.06176	0.03028	0.00805	0.01450	0.04606
12	2-2-3	0.11025	0.10398	0.01434	0.07587	0.13209

Note that since the INIT= option was omitted from the PROC HAPLOTYPE statement, the initial haplotype frequencies used in the EM algorithm are identical to the frequencies that appear in the H0 FREQ column in Output 5.3.1. The frequencies in the H1 FREQ column are those calculated from the final iteration of the EM algorithm, and these frequencies' standard errors and confidence limits are included in the table as well.

Output 5.3.2. Testing for Linkage Disequilibrium Using the LD Option

Test for Allelic Associations

Hypothesis	DF	LogLike	Chi-Square	Pr > ChiSq
H0: No Association	4	-953.89697		
H1: Allelic Associations	11	-934.98180	37.8303	<.0001

Output 5.3.2 displays the log likelihood under the null hypothesis assuming independence among all the loci and the alternative, which allows for associations between markers. The empirical chi-square test statistic of the likelihood ratio test is calculated as $X^2 = 2[-934.98180 - (-953.89697)] = 37.8303$ with degrees of freedom $\nu = 11 - 4 = 7$ that gives a p-value < 0.0001. The test indicates significant linkage disequilibrium among the three loci, as shown in the online documentation from the Lab of Statistical Genetics at Rockefeller University (2001).

Example 5.4. Testing for Marker-Trait Associations

To demonstrate how the TRAIT statement can be utilized, a subset of data from GAW12 (Wijsman et al. 2001) is read into a SAS data set as follows:

```
data gaw;
   input status $ a1-a24;
   datalines;
U 8 4 4 4 2 7 3 2 1 4 10 2  6 6 1 2 1 1 7  7  8 7 8  8
U 5 9 3 5 3 4 2 3 4 3 14 10 3 6 7 7 1 4 5  12 3 3 1  2
A 8 2 5 1 6 3 3 5 3 4 5  3  3 1 5 3 3 4 7  7  7 3 7  7
U 7 8 5 3 8 4 5 3 3 4 13 8  1 3 4 5 4 4 10 7  1 2 2  2
U 9 2 2 5 7 6 9 3 2 4 3  2  5 2 1 2 2 4 5  7  4 3 1  12
U 2 7 1 4 6 7 8 4 4 3 10 5  5 2 4 3 3 1 8  11 2 3 7  7
U 7 7 6 6 1 4 9 5 3 1 14 6  5 3 1 3 3 1 12 1  3 7 7  7
U 4 4 3 7 3 2 8 9 3 1 9  10 6 4 5 3 1 4 10 8  8 5 8  2
A 8 9 6 5 6 4 3 4 4 1 9  1  7 7 2 5 4 1 1  1  5 1 10 2
U 9 5 6 1 2 6 3 3 3 2 8  7  1 5 3 8 1 3 1  8  3 5 1  4
U 8 1 1 5 8 6 3 3 4 3 1  10 3 1 2 3 4 4 5  10 4 5 7  9
A 7 2 3 4 1 3 2 3 3 3 7  1  7 7 2 3 3 4 5  1  5 5 7  9
U 9 3 1 1 2 3 9 8 3 1 13 13 7 1 2 2 3 4 10 3  1 1 10 1
U 2 9 6 1 3 4 3 2 4 3 2  1  4 3 8 1 4 3 9  5  4 2 1  10
U 2 1 1 4 4 7 5 8 3 4 10 13 5 4 4 4 4 3 12 2  3 7 2  12
U 7 7 6 6 3 3 9 3 4 3 14 14 2 1 2 2 1 4 9  1  5 8 4  10
U 1 3 6 5 5 4 9 4 3 4 13 1  2 3 1 2 1 3 1  3  5 3 2  1
U 9 2 6 6 3 4 3 4 2 4 14 9  5 2 4 4 1 1 12 7  5 5 11 7
U 3 3 5 5 8 4 6 5 4 3 2  13 7 1 1 2 3 2 10 7  3 4 7  10
U 4 3 4 5 7 7 8 8 3 3 8  13 3 4 3 2 4 1 1  12 1 3 10 7
U 3 8 1 1 3 8 8 3 4 4 13 12 1 4 5 7 1 4 1  8  3 2 3  3
U 7 8 5 7 7 3 3 3 4 3 14 5  5 1 8 5 4 4 12 12 5 5 10 10
A 7 2 5 4 1 3 3 9 4 3 13 9  2 3 6 5 4 4 1  10 5 2 1  10
U 7 2 4 5 6 1 1 2 4 4 10 8  4 5 5 4 1 1 6  9  2 7 2  12
U 3 3 4 2 7 3 8 3 4 4 14 12 3 2 5 4 3 3 9  3  2 1 12 12
A 2 3 4 1 4 3 3 3 4 4 6  14 1 1 2 2 1 3 3  1  2 8 2  7
U 5 9 3 1 7 4 3 4 2 4 9  8  5 7 3 1 1 3 9  9  2 5 1  9
U 8 5 6 5 3 7 4 4 4 3 10 9  7 5 2 8 4 1 7  8  2 7 12 1
U 9 8 5 5 7 3 6 5 1 3 13 5  2 2 8 7 3 3 9  12 1 3 4  1
A 7 8 5 2 3 5 3 9 3 3 12 5  1 1 1 2 1 4 7  2  5 3 6  1
A 5 4 1 1 3 7 4 5 3 3 14 13 7 3 3 1 4 3 1  8  3 3 2  9
U 8 9 3 2 7 3 8 9 4 1 1  12 5 4 4 6 3 4 2  7  5 2 3  10
A 9 2 3 5 3 3 2 3 2 3 14 13 6 1 3 1 4 3 3  2  3 1 1  7
A 2 5 7 5 6 7 9 4 3 4 14 13 5 1 2 3 4 4 2  10 3 1 12 12
U 7 2 3 1 1 3 4 4 3 4 2  8  5 3 4 6 3 3 10 12 8 3 2  1
A 7 5 1 5 3 3 9 2 3 3 10 6  1 7 2 4 4 4 10 9  1 8 7  3
U 3 2 5 5 4 3 3 5 1 3 1  1  5 2 1 2 3 3 10 3  3 3 10 4
A 3 2 5 5 8 5 3 7 4 3 2  14 5 5 3 3 3 4 11 1  6 2 1  10
A 2 7 5 5 3 2 9 4 3 3 1  7  7 5 4 7 4 1 12 7  2 3 12 9
A 5 7 2 3 7 3 3 3 3 4 9  2  4 1 2 7 1 4 6  1  2 1 7  7
U 7 4 3 4 5 3 3 8 3 3 2  8  4 6 7 7 4 1 3  1  2 4 12 1
U 7 8 5 4 4 7 9 9 4 3 5  13 7 1 4 4 4 4 9  8  8 3 3  10
U 2 8 4 5 3 7 3 4 3 3 8  14 6 4 6 2 3 4 7  1  3 3 3  10
U 6 8 1 3 6 7 5 4 3 4 1  12 3 7 8 4 3 4 12 12 4 7 12 6
A 8 7 3 1 3 6 4 4 3 3 4  10 6 5 8 1 1 4 1  10 2 2 5  2
```

```
U 2 8 6 6 4 8 4 3 4 3 9  1  1 1 2 3 4 4 2  6  2 3 9  7
U 9 8 4 3 7 3 8 4 4 3 8  8  6 6 4 5 3 4 5  5  1 8 10 1
U 9 3 5 1 8 6 5 3 3 2 13 2  3 5 8 2 1 3 1  10 3 3 10 12
U 2 9 1 6 7 4 9 9 4 1 8  1  3 2 5 8 4 4 3  1  3 3 12 7
U 8 8 6 2 3 2 2 4 3 4 6  12 3 1 7 2 4 4 5  9  2 3 1  10
;
```

This data set contains twelve markers. Suppose you are interested in testing three of the marker loci at a time for association with the trait (status in this case: "A" for affected or "U" for unaffected with a particular disease) over all of their haplotypes. That is, assuming the markers are numbered in the order they appear on the chromosome, haplotypes at marker loci 1 through 3 are analyzed, then haplotypes at marker loci 4 through 6 are analyzed, and so on. These tests may be performed in addition to, or in place of, single-marker case-control tests (see Chapter 3 for more information). In order to reduce the amount of SAS code needed for this analysis, a SAS macro can be used as follows:

```
%macro hap_trait;
 %do firsta=1 %to 19 %by 6;
  %let lasta=%eval(&firsta+5);
  %let firstm=%eval((&firsta+1)/2);
  %let lastm=%eval(&lasta/2);
  title "Markers &firstm through &lastm";

  proc haplotype data=gaw noprint;
     var a&firsta-a&lasta;
     trait status;
  run;

 %end;
%mend;
%hap_trait
```

Since the NOPRINT option is specified, this code produces only the "Test for Marker-Trait Association" table each of the four times PROC HAPLOTYPE is invoked.

Output 5.4.1. Testing for Marker-Trait Associations Using Haplotypes

Markers 1 through 3

The HAPLOTYPE Procedure

Test for Marker-Trait Association

Trait Number	Trait Value	Num Obs	DF	LogLike	Chi-Square	Pr > ChiSq
1	U	36	156	-245.18487		
2	A	14	68	-69.90500		
	Combined	50	181	-355.16139	40.0715	0.5990

Markers 4 through 6

The HAPLOTYPE Procedure

Test for Marker-Trait Association

Trait Number	Trait Value	Num Obs	DF	LogLike	Chi-Square	Pr > ChiSq
1	U	36	140	-236.78471		
2	A	14	62	-78.22280		
	Combined	50	162	-349.30084	34.2933	0.7243

Markers 7 through 9

The HAPLOTYPE Procedure

Test for Marker-Trait Association

Trait Number	Trait Value	Num Obs	DF	LogLike	Chi-Square	Pr > ChiSq
1	U	36	119	-242.53993		
2	A	14	56	-68.34854		
	Combined	50	139	-348.95917	38.0707	0.3753

Markers 10 through 12

The HAPLOTYPE Procedure

Test for Marker-Trait Association

Trait Number	Trait Value	Num Obs	DF	LogLike	Chi-Square	Pr > ChiSq
1	U	36	180	-268.92245		
2	A	14	75	-85.15400		
	Combined	50	233	-395.70275	41.6263	0.0069

Output 5.4.1 displays the four tables that are created by this macro. The first corresponds to testing the three-locus haplotypes at the first three marker loci with the TRAIT variable, the second to the second set of three markers, and so on. From the LRTs that are performed and summarized in the output, it can be concluded that out of the four sets of marker loci tested, only haplotypes at markers 10, 11, and 12 show a significant association with the trait variable status. The chi-square statistic for testing the haplotypes at these markers for association with disease status is calculated as $41.62630 = 2(-268.9224489 - 85.15400274 + 395.70275165)$ with degrees of freedom $22 = 180 + 75 - 233$, which has a p-value of 0.0069.

Suppose you would like to further explore the association between these three markers and the trait. You can also perform tests of association between each individual haplotype at these marker loci and disease status using the following code:

```
ods output haplotype.haptraittest=outhap;
proc haplotype data=gaw noprint;
   var a19-a24;
   trait status / testall perm=100;
run;

proc print data=outhap(obs=20) noobs;
   title 'The HAPLOTYPE Procedure';
   title2 ' ';
   title3 'Tests for Haplotype-Trait Association';
run;
ods output close;
```

The TESTALL option indicates that a test for trait association should be performed on each haplotype using a chi-square test statistic, which is performed by default. In addition, since the PERM=100 option is included, an empirical p-value is calculated. Due to the number of alleles at each marker in this example, this option increases the computation time substantially, even with this small number of permutations.

Output 5.4.2. Using the TESTALL Option on Markers 10-12

The HAPLOTYPE Procedure

Tests for Haplotype-Trait Association

Number	Haplotype	Trait1Freq	Trait2Freq	Combined Freq	ChiSq	Prob ChiSq	Prob Exact
1	1-1-2	0.00000	0.03571	0.00000	0	1.0000	1.0000
2	1-1-7	0.00000	0.00000	0.01000	1.0101	0.3149	0.3600
3	1-1-10	0.00000	0.00000	0.01950	1.9883	0.1585	0.1800
4	1-2-1	0.00000	0.01786	0.03000	2.3686	0.1238	0.1700
5	1-2-2	0.00000	0.05357	0.01000	6.0967	0.0135	0.0600
6	1-2-3	0.00000	0.00000	0.00000	0.001666	0.9674	0.6100
7	1-2-5	0.00000	0.00000	0.01000	1.0101	0.3149	0.2700
8	1-2-7	0.00000	0.05357	0.00000	0	1.0000	1.0000
9	1-2-10	0.00000	0.01786	0.00000	0	1.0000	1.0000
10	1-2-12	0.00000	0.00000	0.00000	0	1.0000	1.0000
11	1-3-1	0.00694	0.00000	0.00000	0	1.0000	1.0000
12	1-3-2	0.00000	0.01786	0.01000	0.9019	0.3423	0.4300
13	1-3-3	0.02777	0.00000	0.02000	0.7934	0.3731	0.7500
14	1-3-4	0.00000	0.00000	0.00000	0	1.0000	1.0000
15	1-3-7	0.04167	0.00000	0.02045	2.2035	0.1377	0.1700
16	1-3-9	0.00000	0.01786	0.00000	0	1.0000	1.0000
17	1-3-10	0.00000	0.00000	0.00000	7.8011E-8	0.9998	0.9300
18	1-3-12	0.01389	0.00000	0.01006	0.3905	0.5320	0.8200
19	1-4-1	0.01389	0.00000	0.00000	0	1.0000	1.0000
20	1-4-12	0.00000	0.00000	0.00000	0	1.0000	1.0000

Output 5.4.2 displays the table "Test for Haplotype-Trait Association" as a SAS data set using the ODS system in order to show only the first 20 rows. The table contains haplotypes at markers 10, 11, and 12 and their estimated frequencies among individuals with the first trait value, individuals with the second trait value, and all individuals. The chi-square statistic testing whether the frequencies between the two trait groups are significantly different is also shown, along with its 1 df p-value. Note that none of the haplotypes shown here have a significant association with disease status.

Example 5.5. Creating a Data Set for Haplotype Trend Regression (HTR)

Zaykin et al. (2002) discuss HTR, an approach to testing haplotypes for association with a phenotype using a regression model, which can be more powerful than the omnibus chi-square test performed in PROC HAPLOTYPE. The output data set produced by PROC HAPLOTYPE can easily be transformed into one that can be used by one of the regression procedures offered by SAS/STAT to implement HTR.

Here is an example data set that can be analyzed using PROC HAPLOTYPE:

```
data alleles;
   input (a1-a6) ($) disease;
   datalines;
A  a  B  B  c  C  1
A  A  B  b  c  C  1
a  A  B  b  c  c  0
A  A  B  B  c  C  1
A  A  b  B  c  C  1
A  A  B  b  C  c  0
A  a  b  B  C  c  1
A  A  b  B  C  c  1
A  a  B  B  c  c  1
a  a  B  b  c  c  0
A  A  B  B  C  C  1
A  A  B  B  c  c  1
a  A  b  b  c  c  0
A  A  B  B  c  c  1
A  A  b  b  c  c  0
A  A  b  B  c  C  0
A  A  B  b  c  C  1
A  a  b  B  c  c  1
A  a  B  B  c  C  1
A  A  b  b  C  C  0
A  A  B  B  C  C  1
A  A  b  B  C  c  1
A  A  b  B  c  C  1
a  A  B  b  C  c  0
A  a  B  B  C  C  0
A  A  B  B  C  c  1
A  A  B  b  C  c  0
A  A  B  B  c  C  1
a  A  B  b  C  C  1
A  a  B  b  C  c  1
A  A  B  b  c  C  1
A  a  B  B  c  c  1
A  A  B  b  C  c  1
a  A  B  b  C  c  1
A  A  B  b  C  C  1
A  a  B  B  C  C  1
a  A  B  b  C  c  0
a  A  b  B  C  C  0
A  A  B  b  c  C  1
a  A  B  b  c  c  0
A  A  B  B  C  C  0
A  A  B  B  c  c  1
A  a  B  B  C  c  1
;
```

An output data set containing individuals' probabilities of having particular haplotype pairs can be created in the usual manner, while also performing an omnibus test for association between the three markers and disease status.

```
proc haplotype data=alleles out=out;
   var a1-a6;
   trait disease;
run;
```

This code executes the omnibus marker-trait association test whose p-value is given by the chi-square distribution.

Output 5.5.1. Testing for an Overall Marker-Trait Association

The HAPLOTYPE Procedure

Test for Marker-Trait Association

Trait Number	Trait Value	Num Obs	DF	LogLike	Chi-Square	Pr > ChiSq
1	1	29	7	-68.11558		
2	0	14	7	-37.28544		
	Combined	43	7	-115.48338	10.0824	0.1840

Output 5.5.1 shows that there is no significant association between the markers and the trait, disease status. However, the more powerful HTR can be used to perform the same test.

```
data out1;
   set out;
   haplotype=tranwrd(haplotype1,'-','_');

data out2;
   set out;
   haplotype=tranwrd(haplotype2,'-','_');

data outnew;
   set out1 out2;

proc sort data=outnew;
   by haplotype;
run;

data outnew2;
   set outnew;
   lagh=lag(haplotype);
   if haplotype ne lagh then num+1;
   hapname=compress("H"||num,' ');

proc sort data=outnew2;
   by id hapname;
run;

data outt;
   set outnew2;
   by id haplotype;
   if first.haplotype then totprob=prob/2;
   else totprob+prob/2;
   if last.haplotype;

proc transpose data=outt out=outreg(drop=_NAME_) ;
   id hapname;
   idlabel haplotype;
   var totprob;
   by id;
run;

data outmissto0(drop=i);
   set outreg;
   array h{8};
   do i=1 to 8;
    if h{i}=. then h{i}=0;
   end;

data htr(keep=id disease h1-h8);
   merge outmissto0 alleles;
run;

proc print data=htr noobs round label;
run;
```

```
proc logistic data=all descending;
   model disease = h1-h8 / selection=stepwise;
run;
```

This SAS code produces a data set htr from the output data set of PROC HAPLOTYPE that contains the variables needed to be able to perform HTR. There is now one column for each possible haplotype in the sample, with each column containing the haplotype's frequency, or probability, within an individual.

Output 5.5.2. HTR Data Set

ID	A_B_C	A_B_c	a_B_C	a_B_c	A_b_C	A_b_c	a_b_c	a_b_C	disease
1	0.29	0.21	0.21	0.29	0.00	0.00	0.00	0.00	1
2	0.27	0.23	0.00	0.00	0.23	0.27	0.00	0.00	1
3	0.00	0.27	0.00	0.23	0.00	0.23	0.27	0.00	0
4	0.50	0.50	0.00	0.00	0.00	0.00	0.00	0.00	1
5	0.27	0.23	0.00	0.00	0.23	0.27	0.00	0.00	1
6	0.27	0.23	0.00	0.00	0.23	0.27	0.00	0.00	0
7	0.22	0.00	0.13	0.15	0.15	0.13	0.22	0.00	1
8	0.27	0.23	0.00	0.00	0.23	0.27	0.00	0.00	1
9	0.00	0.50	0.00	0.50	0.00	0.00	0.00	0.00	1
10	0.00	0.00	0.00	0.50	0.00	0.00	0.50	0.00	0
11	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1
12	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1
13	0.00	0.00	0.00	0.00	0.00	0.50	0.50	0.00	0
14	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1
15	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0
16	0.27	0.23	0.00	0.00	0.23	0.27	0.00	0.00	0
17	0.27	0.23	0.00	0.00	0.23	0.27	0.00	0.00	1
18	0.00	0.27	0.00	0.23	0.00	0.23	0.27	0.00	1
19	0.29	0.21	0.21	0.29	0.00	0.00	0.00	0.00	1
20	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0
21	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1
22	0.27	0.23	0.00	0.00	0.23	0.27	0.00	0.00	1
23	0.27	0.23	0.00	0.00	0.23	0.27	0.00	0.00	1
24	0.22	0.00	0.13	0.15	0.15	0.13	0.22	0.00	0
25	0.50	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0
26	0.50	0.50	0.00	0.00	0.00	0.00	0.00	0.00	1
27	0.27	0.23	0.00	0.00	0.23	0.27	0.00	0.00	0
28	0.50	0.50	0.00	0.00	0.00	0.00	0.00	0.00	1
29	0.01	0.00	0.49	0.00	0.49	0.00	0.00	0.01	1
30	0.22	0.00	0.13	0.15	0.15	0.13	0.22	0.00	1
31	0.27	0.23	0.00	0.00	0.23	0.27	0.00	0.00	1
32	0.00	0.50	0.00	0.50	0.00	0.00	0.00	0.00	1
33	0.27	0.23	0.00	0.00	0.23	0.27	0.00	0.00	1
34	0.22	0.00	0.13	0.15	0.15	0.13	0.22	0.00	1
35	0.50	0.00	0.00	0.00	0.50	0.00	0.00	0.00	1
36	0.50	0.00	0.50	0.00	0.00	0.00	0.00	0.00	1
37	0.22	0.00	0.13	0.15	0.15	0.13	0.22	0.00	0
38	0.01	0.00	0.49	0.00	0.49	0.00	0.00	0.01	0
39	0.27	0.23	0.00	0.00	0.23	0.27	0.00	0.00	1
40	0.00	0.27	0.00	0.23	0.00	0.23	0.27	0.00	0
41	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0
42	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	1
43	0.29	0.21	0.21	0.29	0.00	0.00	0.00	0.00	1

The data set shown in Output 5.5.2 can now be used in one of the regression procedures offered by SAS/STAT. In this example, since the trait is binary, the LOGISTIC procedure can be used to perform HTR on the variable disease. Since HTR can be more powerful than the chi-square test based on a contingency table, it might find a significant association between the markers in the data set and disease status that the chi-square LRT did not detect.

Output 5.5.3. PROC LOGISTIC Output

The LOGISTIC Procedure

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	6.1962	1	0.0128
Score	6.3995	1	0.0114
Wald	4.9675	1	0.0258

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.1986	0.4058	8.7224	0.0031
H8	1	-6.3249	2.8378	4.9675	0.0258

Output 5.5.3 shows two of the tables produced by PROC LOGISTIC. The first one displays the test of the global null hypothesis, $\beta = \mathbf{0}$. You can see that this test indicates a significant association between the haplotypes at the three markers and disease status. In particular, the second table shows that as a result of the stepwise selection, the haplotype H8 (a-b-c) has a statistically significant effect on disease status. This is an example of a case where HTR detects a significant association that the LRT used in PROC HAPLOTYPE does not.

References

Excoffier, L. and Slatkin, M. (1995), "Maximum-Likelihood Estimation of Molecular Haplotype Frequencies in a Diploid Population," *Molecular Biology and Evolution*, 12, 921–927.

Fallin, D., Cohen, A., Essioux, L., Chumakov, I., Blumenfeld, M., Cohen, D., and Schork, N.J. (2001), "Genetic Analysis of Case/Control Data Using Estimated Haplotype Frequencies: Application to APOE Locus Variation and Alzheimer's Disease," *Genome Research*, 11, 143–151.

Fallin, D. and Schork, N.J. (2000), "Accuracy of Haplotype Frequency Estimation for Biallelic Loci, via the Expectation-Maximization Algorithm for Unphased Diploid Genotype Data," *American Journal of Human Genetics*, 67, 947–959.

Hawley, M.E. and Kidd, K.K. (1995), "HAPLO: A Program Using the EM Algorithm to Estimate the Frequencies of Multi-site Haplotypes," *Journal of Heredity*, 86, 409–411.

Lab of Statistical Genetics at Rockefeller University (2001), "User's Guide to the EH Program," [http://linkage.rockefeller.edu/ott/eh.htm].

Long, J.C., Williams, R.C., and Urbanek, M. (1995), "An E-M Algorithm and Testing Strategy for Multiple-Locus Haplotypes," *American Journal of Human Genetics*, 56, 799–810.

Wijsman, E.M., Almasy, L., Amos, C.I., Borecki, I., Falk, C.T., King, T.M., Martinez, M.M., Meyers, D., Neuman, R., Olson, J.M., Rich, S., Spence, M.A., Thomas, D.C., Vieland, V.J., Witte, J.S., and MacCluer, J.W. (2001), "Analysis of Complex Genetic Traits: Applications to Asthma and Simulated Data," *Genetic Epidemiology*, 21, S1–S853.

Zaykin, D.V., Westfall, P.H., Young, S.S., Karnoub, M.A., Wagner, M.J., and Ehm, M.G. (2002), "Testing Association of Statistically Inferred Haplotypes with Discrete and Continuous Traits in Samples of Unrelated Individuals," to appear in *Human Heredity*.

Zhao, J.H., Curtis, D., and Sham, P.C. (2000), "Model-Free Analysis and Permutation Tests for Allelic Associations," *Human Heredity*, 50, 133–139.

Chapter 6
The PSMOOTH Procedure

Chapter Contents

Chapter 6
The PSMOOTH Procedure

Overview

In the search for complex disease genes, linkage and/or association tests are often performed on markers from a genome-wide scan or SNPs from a finely scaled map. This means hundreds or even thousands of hypotheses are being simultaneously tested. Plotting the negative log p-values of all the marker tests will reveal many peaks that indicate significant test results, some of which are false positives. In order to reduce the number of false positives or improve power, smoothing methods can be applied that take into account p-values from neighboring, and possibly correlated, markers. That is, the peak length can be used to indicate significance in addition to the peak height. The PSMOOTH procedure offers smoothing methods that implement either Simes' (1986) or Fisher's (1932) method for multiple hypothesis testing. These methods 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 available as well.

PROC PSMOOTH can take any data set containing any number of columns of p-values as an input data set, including the output data sets from the CASECONTROL and FAMILY procedures (see Chapter 3 and Chapter 4 for more information).

Getting Started

Example

Suppose you have tested 48 markers for association with a disease using the genotype case-control and trend tests in PROC CASECONTROL. Now you are concerned about the multiple hypothesis testing issue, and so you run PROC PSMOOTH on the output data set from PROC CASECONTROL in order to eliminate the number of false positives found using the individual p-values from the marker-trait association tests.

```
proc casecontrol data=in outstat=cc_tests genotype trend;
   trait affected;
   var a1-a96;
run;

proc psmooth data=cc_tests simes fisher bw=2 sidak out=adj_p;
   var ProbGenotype ProbTrend;
   id Locus;
run;

proc print data=adj_p;
run;
```

This code modifies the p-values contained in the output data set from PROC CASECONTROL, first by smoothing the p-values using both Simes' and Fisher's methods with a bandwidth of 2, then by applying Sidak's multiple testing adjustment to the smoothed p-values.

Obs	Locus	Prob Genotype	Prob Genotype_ S2	Prob Genotype_ F2	Prob Trend	Prob Trend_S2	Prob Trend_F2
1	M01	0.20059	1.00000	1.00000	0.12947	1.00000	1.00000
2	M02	0.24984	1.00000	1.00000	0.27988	1.00000	1.00000
3	M03	0.61303	0.99989	0.99559	0.64512	0.99990	0.99176
4	M04	0.35724	0.99989	0.99992	0.38009	0.99990	0.99999
5	M05	0.03454	0.99989	1.00000	0.03486	0.99990	1.00000
6	M06	0.51121	0.99989	1.00000	0.57352	0.99990	1.00000
7	M07	0.48462	0.99989	0.99996	0.59738	0.99990	1.00000
8	M08	0.56565	1.00000	1.00000	0.66992	1.00000	1.00000
9	M09	0.22166	1.00000	1.00000	0.19083	1.00000	1.00000
10	M10	0.56964	1.00000	1.00000	0.68374	1.00000	1.00000
11	M11	0.42669	1.00000	1.00000	0.48177	1.00000	1.00000
12	M12	0.34151	1.00000	1.00000	0.39099	1.00000	1.00000
13	M13	0.35094	1.00000	1.00000	0.33305	1.00000	1.00000
14	M14	0.41437	1.00000	1.00000	0.44417	1.00000	1.00000
15	M15	0.36395	1.00000	1.00000	0.40781	1.00000	1.00000
16	M16	0.45854	1.00000	1.00000	0.50985	1.00000	1.00000
17	M17	0.61439	1.00000	1.00000	0.65185	1.00000	1.00000
18	M18	0.35993	1.00000	1.00000	0.41276	1.00000	1.00000
19	M19	0.26623	1.00000	1.00000	0.27373	1.00000	1.00000
20	M20	0.39734	0.93037	0.85928	0.47567	0.99978	0.97505
21	M21	0.18037	0.93037	0.79634	0.11657	0.99978	0.91010
22	M22	0.01080	0.93037	0.92348	0.03222	0.99978	0.98274
23	M23	0.26079	0.93037	0.43744	0.20915	0.98714	0.52502
24	M24	0.56114	0.93037	0.57991	0.63033	0.98714	0.79232
25	M25	0.04130	0.99998	0.99312	0.03468	0.99989	0.97502
26	M26	0.32966	0.99998	0.99112	0.35792	0.99989	0.98829
27	M27	0.16512	0.99998	0.95182	0.12327	0.99989	0.94535
28	M28	0.23915	1.00000	0.99999	0.28337	1.00000	1.00000
29	M29	0.27391	1.00000	0.99996	0.31744	1.00000	0.99999
30	M30	0.39576	1.00000	1.00000	0.40239	1.00000	1.00000
31	M31	0.25123	1.00000	0.99970	0.30787	0.98124	0.98142
32	M32	0.32687	1.00000	1.00000	0.36567	0.98124	0.99795
33	M33	0.08401	1.00000	1.00000	0.01590	0.98124	0.99975
34	M34	0.56936	0.99999	0.99962	0.65595	0.98124	0.96888
35	M35	0.60157	0.99999	0.99774	0.66286	0.98124	0.89811
36	M36	0.07535	1.00000	0.99999	0.07229	1.00000	1.00000
37	M37	0.21207	1.00000	0.99937	0.18823	1.00000	0.99950
38	M38	0.24277	0.83623	0.35131	0.29450	0.90310	0.39211
39	M39	0.27190	0.80769	0.11928	0.25095	0.90310	0.22619
40	M40	0.00740	0.80769	0.02802	0.00949	0.86357	0.04171
41	M41	0.01351	0.80769	0.01060	0.02809	0.83848	0.00730
42	M42	0.02846	0.80769	0.01860	0.01626	0.83848	0.01510
43	M43	0.06777	0.96512	0.34054	0.02982	0.91341	0.25411
44	M44	0.56635	0.99937	0.97025	0.64220	0.97572	0.83276
45	M45	0.50034	1.00000	1.00000	0.51227	0.99957	0.99989
46	M46	0.34136	1.00000	1.00000	0.40320	1.00000	1.00000
47	M47	0.25543	1.00000	0.99996	0.23193	1.00000	0.99998
48	M48	0.08614	1.00000	0.99900	0.08509	1.00000	0.99926

Figure 6.1. PROC PSMOOTH Output Data Set

Figure 6.1 displays the original and modified p-values.

Syntax

The following statements are available in PROC PSMOOTH.

PROC PSMOOTH < *options* > **;**
BY *variables* **;**
ID *variables* **;**
VAR *variables* **;**

Items within angle brackets (< >) are optional, and statements following the PROC PSMOOTH statement can appear in any order. The VAR statement is required. The syntax of each statement is described in the following section in alphabetical order after the description of the PROC PSMOOTH statement.

PROC PSMOOTH Statement

PROC PSMOOTH < *options* > **;**

You can specify the following options in the PROC PSMOOTH statement.

BANDWIDTH=*number list*
BW=*number list*

gives the values for the bandwidths to use in combining p-values. A bandwidth of w indicates that w p-values on each side of the original p-value are included in the combining method to create a sliding window of size $2w+1$. The number list can contain any combination of the following forms, with each form separated by a comma:

$w_1, w_2, ..., w_n$	a list of several values
w_1 to w_2	a sequence where w_1 is the starting value, w_2 is the ending value, and the increment is 1.
w_1 to w_2 by i	a sequence where w_1 is the starting value, w_2 is the ending value, and the increment is i.

All numbers in the number list must be integers, and any negative numbers are ignored. An example of a valid number list is

```
bandwidth = 1,2, 5 to 15 by 5, 18
```

which would perform the combining of p-values using bandwidths 1, 2, 5, 10, 15, and 18, which create sliding windows of size 3, 5, 11, 21, 31, and 37, respectively.

BONFERRONI
BON

requests that the Bonferroni adjustment for multiple testing based on the number of observations in the BY group be applied to the p-values in the output data set. This adjustment is applied after the smoothing has occurred. This option is ignored if the SIDAK option is specified.

DATA=*SAS-data-set*
names the input SAS data set to be used by PROC PSMOOTH. If this option is omitted, the SAS system option _LAST_ is used, which by default is the most recently created data set.

FISHER
requests that Fisher's method for combining p-values from multiple hypotheses be applied to the original p-values.

NEGLOG
requests that all p-values, original and combined, be transformed to their negative log (base e) in the output data set; that is, for each p-value, $-\ln(p\text{-value})$ is reported in the OUT= data set. This option is useful for graphing purposes.

NEGLOG10
requests that all p-values, original and combined, be transformed to their negative log (base 10) in the output data set; that is, for each p-value, $-\log_{10}(p\text{-value})$ is reported in the OUT= data set. This option is useful for graphing purposes.

OUT=*SAS-data-set*
names the output SAS data set containing the original p-values and the new combined p-values. When this option is omitted, an output data set is created by default and named according to the DATA*n* convention.

SIDAK
requests that the Sidak adjustment for multiple testing based on the number of observations in the BY group be applied to the p-values in the output data set. This adjustment is applied after the smoothing has occurred.

SIMES
requests that Simes' method for combining p-values from multiple hypotheses be applied to the original p-values.

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC PSMOOTH to obtain separate analyses on observations in groups defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables. The *variables* are one or more variables in the input data set.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data using the SORT procedure with a similar BY statement.
- Specify the BY statement option NOTSORTED or DESCENDING in the BY statement for the PSMOOTH procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables using the DATASETS procedure (in Base SAS software).

For more information on the BY statement, refer to the discussion in *SAS Language Reference: Concepts*. For more information on the DATASETS procedure, refer to the discussion in the *SAS Procedures Guide*.

ID Statement

ID *variables* **;**

The ID statement identifies the variables from the DATA= data set that should be included in the OUT= data set.

VAR Statement

VAR *variables* **;**

The VAR statement identifies the variables containing the original p-values on which the combining methods should be performed.

Details

Statistical Computations

Methods for Smoothing p-Values

PROC PSMOOTH offers two methods for combining p-values over specified sizes of sliding windows. For each value w listed in the BANDWIDTH= option of the PROC PSMOOTH statement, a sliding window of size $2w + 1$ is used; that is, the p-values for each set of $2w + 1$ consecutive markers are considered in turn, for each value w. The approach described by Zaykin et al. (2002) is implemented, where the original p-value at the center of the sliding window is replaced by a function of the original p-value and the p-values from the w nearest markers on each side to create a new sequence of p-values. Note that for markers less than w from the beginning or end of the data set (or BY group if any variables are specified in the BY statement), the number of hypotheses tested, L, is adjusted accordingly. The two methods for combining p-values from multiple hypotheses are Simes' method and Fisher's method, described in the following two sections. Plotting the new p-values versus the original p-values reveals the smoothing effect this technique has.

Simes' Method

Simes' method for combining p-values is performed as follows when the SIMES option is specified in the PROC PSMOOTH statement: let p_j be the original p-value at the center of the current sliding window, which contains $p_{j-w}, ..., p_{j+w}$. From these $L = 2w + 1$ p-values, the ordered p-values, $p_{(1)}, ..., p_{(L)}$ are formed. Then the new value for p_j is $\min_{1 \le i \le L}(Lp_{(i)}/i)$.

This method controls the type I error rate even when hypotheses are positively correlated (Sarkar and Chang 1997), which is expected for nearby markers. Thus if dependencies are suspected among tests that are performed, this method is recommended due to its conservativeness.

Fisher's Method

When the FISHER option is issued in the PROC PSMOOTH statement, Fisher's method for combining p-values is applied by replacing the p-value at the center of the current sliding window p_j by the p-value of the statistic t, where

$$t = -2 \sum_{i=j-w}^{j+w} \ln(p_i)$$

which has a χ^2_{2L} distribution under the null hypothesis of all $L = 2w + 1$ hypotheses being true.

CAUTION: t has a χ^2 distribution only under the assumption that the tests performed are mutually independent. When this assumption is violated, the probability of type I error may exceed the significance level α.

Multiple Testing Adjustments for p-Values

While the smoothing methods take into account the p-values from neighboring markers, the number of hypothesis tests performed does not change. Therefore, the Bonferroni and Sidak methods are offered by PROC PSMOOTH to adjust the smoothed p-values for multiple testing. The number of tests performed, R, is the number of observations in the current BY group if any variables are specified in the BY statement, or the number of observations in the entire data set if there are no variables specified in the BY statement. If both the BONFERRONI and SIDAK options are specified in the PROC PSMOOTH statement, only the Sidak method is used. Note that these adjustments will not be applied to the original column(s) of p-values; if you would like to adjust the original p-values for multiple testing, you must include a bandwidth of 0 in the BANDWIDTH= option of the PROC PSMOOTH statement.

For R tests, the p-value p is adjusted as follows according to these two methods:

Bonferroni adjustment: $\min(Rp, 1.0)$

Sidak adjustment (Sidak 1967): $1 - (1 - p)^R$

Both methods are conservative, with Sidak's slightly less conservative than Bonferroni's method.

Missing Values

Missing values in a sliding window, even at the center of the window, are simply ignored, and the number of hypotheses L is reduced accordingly. Thus the smoothing methods can be applied to any window that contains at least one nonmissing value. Any p-values in the input data set that fall outside the interval [0,1] are treated as missing.

OUT= Data Set

The output data set specified in the OUT= option of the PROC PSMOOTH statement contains any BY variables and ID variables. Then for each variable in the VAR statement, the original column is included along with a column for each method and bandwidth specified in the PROC PSMOOTH statement. These variable names are formed by adding the suffixes "_Sw" and "_Fw" for Simes' and Fisher's methods respectively and a bandwidth of size w. For example, if the options BANDWIDTH=1,4 and SIMES and FISHER are all specified in the PROC PSMOOTH statement, and RawP is the variable specified in the VAR statement, the OUT= data set includes RawP, RawP_S1, RawP_F1, RawP_S4, and RawP_F4. If the NEGLOG or NEGLOG10 option is specified in the PROC PSMOOTH statement, then these columns all contain the negative logs (base e or base 10, respectively) of the p-values.

Example

Example 6.1. Displaying Plot of PROC PSMOOTH Output Data Set

Data other than the output data sets from the CASECONTROL and FAMILY procedures can be used in PROC PSMOOTH; here is an example of how to use p-values from another source.

```
data tests;
   input Marker Pvalue @@;
   datalines;
 1  0.72841     2   0.40271
 3  0.32147     4   0.91616
 5  0.27377     6   0.48943
 7  0.40131     8   0.25555
 9  0.57585    10   0.20925
11  0.01531    12   0.23306
13  0.69397    14   0.33040
15  0.97265    16   0.53639
17  0.88397    18   0.03188
19  0.13570    20   0.79138
21  0.99467    22   0.37831
23  0.86459    24   0.97092
25  0.19372    26   0.85339
27  0.32078    28   0.31806
29  0.00655    30   0.82401
31  0.65339    32   0.36115
33  0.92704    34   0.49558
35  0.64842    36   0.43606
37  0.67060    38   0.87520
39  0.78006    40   0.27252
41  0.28561    42   0.80495
43  0.98159    44   0.97030
45  0.53831    46   0.78712
47  0.88493    48   0.36260
49  0.53310    50   0.65709
51  0.26527    52   0.46860
53  0.55465    54   0.54956
55  0.44477    56   0.04933
57  0.12016    58   0.76181
59  0.80158    60   0.18244
61  0.01382    62   0.15100
63  0.04713    64   0.52655
65  0.59368    66   0.94420
67  0.60104    68   0.32848
69  0.90195    70   0.21374
71  0.95471    72   0.14145
73  0.95215    74   0.70330
75  0.19921    76   0.99086
77  0.75736    78   0.23761
79  0.87260    80   0.91472
81  0.33650    82   0.26160
83  0.41948    84   0.62817
85  0.48721    86   0.67093
87  0.53089    88   0.13623
89  0.44344    90   0.41172
;
```

The following code will apply Simes' method for multiple hypothesis testing in order to adjust the p-values.

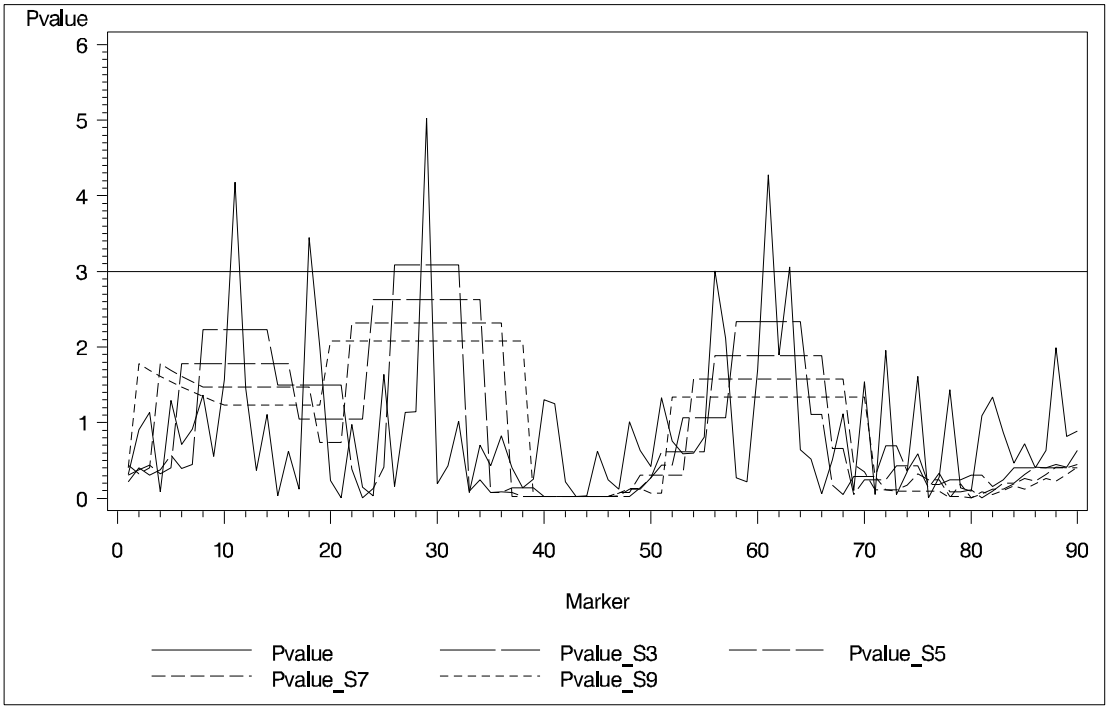
```
proc psmooth data=tests out=pnew simes bandwidth=3 to 9 by 2 neglog;
   var Pvalue;
   id Marker;
run;

symbol1 v=none i=join;
symbol2 v=none i=join line=5;
symbol3 v=none i=join line=4;
symbol4 v=none i=join line=3;
symbol5 v=none i=join line=2;
legend1 label=none;

proc gplot data=pnew;
   plot (Pvalue Pvalue_S3 Pvalue_S5 Pvalue_S7 Pvalue_S9)*Marker
        /overlay vref=3.0 legend=legend1;
run;
```

The NEGLOG option is used in the PROC PSMOOTH statement to facilitate plotting the p-values using the GPLOT procedure of SAS/GRAPH. The plot demonstrates the effect of the different window sizes that are implemented.

Output 6.1.1. Line Plot of Negative Log p-Values



Note how the plots become progressively smoother as the window size increases. Points above the horizontal reference line in Output 6.1.1 represent significant p-values at the 0.05 level. While six of the markers have significant p-values before adjustment, only the method using a bandwidth of 3 finds any significant markers, all in the 26–32 region. This may be an indication that the other five markers are significant only by chance; that is, they may be false positives.

References

Fisher, R.A. (1932), *Statistical Methods for Research Workers*, London: Oliver and Boyd.

Sarkar, S.K. and Chang, C-K. (1997), "The Simes Method for Multiple Hypothesis Testing with Positively Dependent Test Statistics," *Journal of the American Statistical Association*, 92, 1601–1608.

Sidak, Z. (1967), "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions," *Journal of the American Statistical Association*, 62, 626–633.

Simes, R.J. (1986), "An Improved Bonferroni Procedure for Multiple Tests of Significance," *Biometrika*, 73, 751–754.

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.

Chapter 7
The TPLOT Macro

Chapter Contents

Chapter 7
The TPLOT Macro

Overview

The %TPLOT macro creates a triangular plot that graphically displays genetic marker test results. The plot has colors and shapes representing p-value ranges for tests of the following quantities: linkage disequilibrium between pairs of markers, Hardy-Weinberg equilibrium (HWE) for individual markers, and associations between markers and a dichotomous trait (such as disease status). This is a convenient way of combining information contained in output data sets from two separate SAS/Genetics procedures and summarizing it in an easily interpretable plot. Thus, insights can be gleaned by simply studying a plot rather than by having to search through many rows of data or writing code to attempt to summarize the results.

The %TPLOT macro is a part of the SAS Autocall library, and is automatically available for use in your SAS program provided that the SAS system option MAUTOSOURCE is in effect. For more information about autocall libraries, refer to *SAS Macro Language: Reference, Version 8*, 2000.

Syntax

The %TPLOT macro has the following form:

%TPLOT *(SAS-data-set , SAS-data-set , variable [, option])*

The first argument, *SAS-data-set*, specifies the name of the SAS data set that is the output data set from the ALLELE procedure (see Chapter 2), containing the linkage disequilibrium test and HWE test p-values. A user-created data set may be used instead, but is required to contain the variables Locus1 and Locus2 and a variable ProbChi containing the p-values from the disequilibrium tests. The order in which the Locus1 and Locus2 variables are sorted is the order in which the values are displayed on the vertical and horizontal axes, respectively.

The second argument, *SAS-data-set*, specifies the name of the SAS data set that contains the p-values for the marker-trait association tests. This data set can be the output data set from the CASECONTROL procedure, the FAMILY procedure, or the PSMOOTH procedure, or it can be created by the user. A user-created data set must contain a Locus variable for the values on the axes and a variable containing p-values that is specified in the third argument, discussed in the following paragraph. The Locus variable must be in the same sorted order as the Locus1 variable in the data set named in the first argument.

The third argument, *variable*, names the variable that contains the marker-trait association p-values in the SAS data set that is specified in the second argument.

The first three arguments are required. The following option can be used with the %TPLOT macro. The option must follow the three required arguments.

ALPHA= *number*
specifies the significance level for the marker-trait association test. This level is used as a cut-off for the p-value range corresponding to the symbol shape on the plot. This number must be between 0 and 1. The default is ALPHA=0.05.

Results

Plot

Running the %TPLOT macro creates a window displaying a graphical representation of the marker test results.

Here is an example of the TPLOT results window:

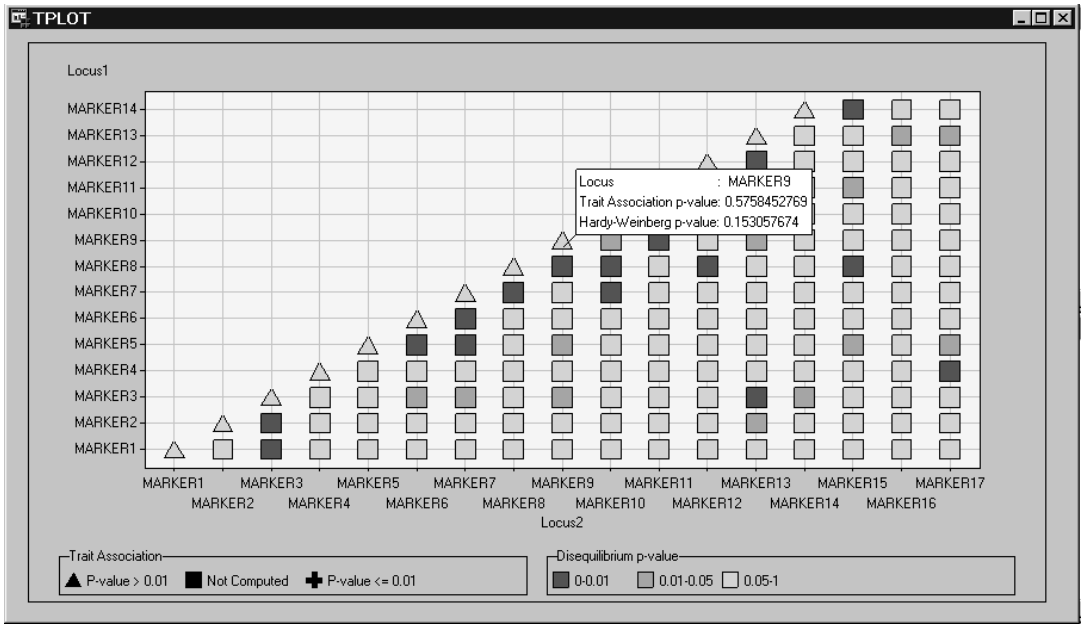


Figure 7.1. Results Window for TPLOT Macro

This plot contains a grid of points with symbols that represent the p-values for various marker tests. Colors and shapes of the data points are used to symbolize p-value ranges. The **Show Info About Points** button in the toolbar enables the p-values to be displayed. While holding down the left-hand mouse button on any point in the plot, the pop-up menu will display for off-diagonal points, the two markers being tested for linkage disequilibrium and the p-value of the test; it displays the marker and its p-values for the HWE test and marker-trait association test for points on the diagonal, as shown in Figure 7.1.

Disequilibrium Tests

The p-values from the linkage disequilibrium tests between all pairs of markers (or all markers within a certain range of each other) are represented by the color of the squares on the off-diagonal of the plot. For the points on the diagonal, the results from the Hardy-Weinberg equilibrium test are displayed instead of the linkage disequilibrium tests since the same marker locus is on the horizontal and vertical axes.

The three ranges of p-values that correspond to different colored symbols in the plot are

Red	[0, 0.01]
Orange	(0.01, 0.05]
Yellow	(0.05, 1]

The disequilibrium test p-values that are plotted can be provided by the output data set from PROC ALLELE, or by a user-created data set meeting the requirements described in the "Syntax" section on page 113.

Marker-Trait Association Tests

Points on the diagonal also display p-values from marker-trait association tests, using the shape of the symbol to correspond to two categories of p-values, significant and not significant. The significance level is set to 0.05 by default, but can be modified using the ALPHA= option in the %TPLOT macro. Thus, for a significance level of α, the following shapes represent the following ranges:

Plus		[0, α]
Triangle		(α, 1]

Note that the square shape  of the off-diagonal points does not represent a marker-trait association p-value since there are two different marker loci represented on the horizontal and vertical axes. These p-values can be provided by the output data set of PROC CASECONTROL, PROC FAMILY, or PROC PSMOOTH. Alternatively, a user-created data set that meets the conditions described in the "Syntax" section (page 113) can be used.

Menu Bar

The results window contains the following pull-down menus:

File

Close	closes the results window.
Print Setup	opens the printer setup utility.
Print	prints the plot as it is currently shown.
Exit	exits the current SAS session.

Edit

Copy	copies the plot to the clipboard.

Format

Rescale Axes when selected, changes the scale of the axes to fit the entire plot in the window.

These menus are also available by clicking the right-hand mouse button anywhere in the TPLOT results window.

Toolbar

A toolbar is displayed at the top of the TPLOT results window. Use the toolbar to display information about points on the plot or to modify the plot's appearance. Tool tips are displayed when you place your mouse pointer over an icon in the toolbar.

Figure 7.2. Toolbar for the %TPLOT Results Window

Tool icons from left to right are as follows:

1. **Print** - prints the plot.
2. **Copy** - copies the plot to the clipboard.
3. **Select a Node or Point** - activates a point on the plot.
4. **Show Info About Points** - displays a text box with information about the selected point.
5. **Scroll Data** - scrolls across data points within the plot. Use this tool when the plot is not able to display all of the points in a single frame.
6. **Move Graph** - moves the plot within the window.
7. **Zoom In/Out** - reduces or increases the size of the plot.
8. **Reset** - returns the plot to its default settings.
9. **What's This?** - displays the help for the results window.

Example

Here is an example of the code that can be used to create the triangular plot of p-values for the data set pop22. This data set is in the proper form for a PROC ALLELE input data set, containing columns of alleles for 150 markers.

```
proc allele data=pop22 outstat=ldstats noprint maxdist=150;
   var a1-a300;
run;

proc casecontrol data=pop22 outstat=assocstats genotype;
   trait affected;
   var a1-a300;
run;

proc psmooth data=assocstats out=sm_assocstats bw=5 simes;
   id Locus;
   var ProbGenotype;
run;

%tplot(ldstats, sm_assocstats, ProbGenotype_S5);
```

Note that the output data set from PROC CASECONTROL can be used in place of the output data set from PROC PSMOOTH if you wish to use unadjusted p-values. This code creates the following plot in the TPLOT window:

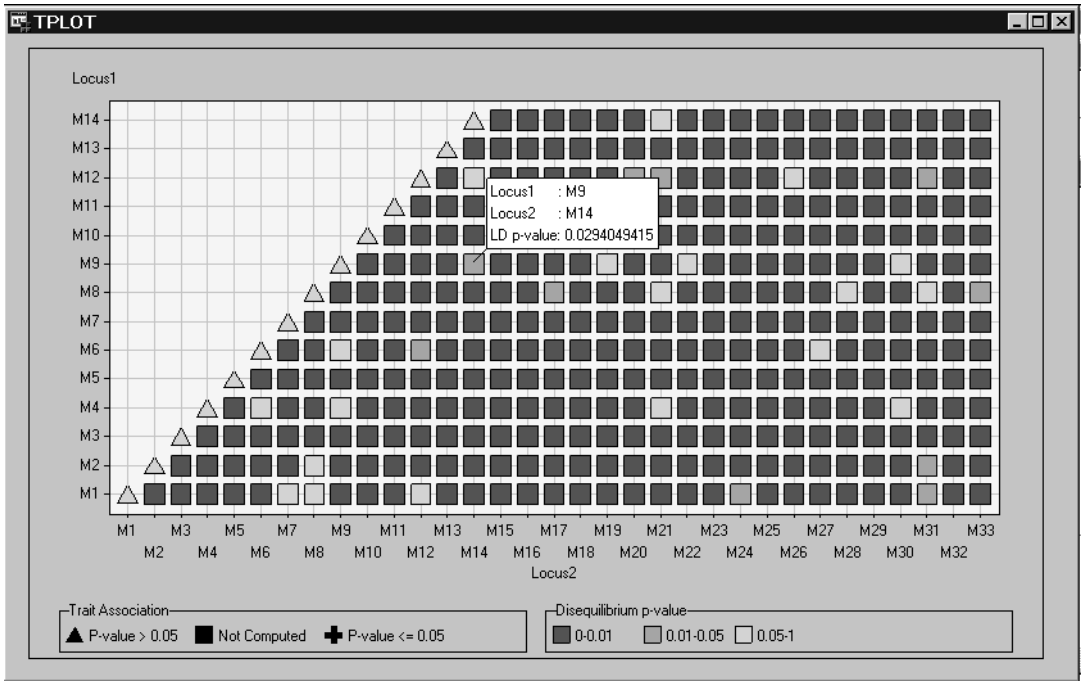


Figure 7.3. Results Window for TPLOT Macro

Figure 7.3 displays the bottom left-hand corner of the plot. The pop-up window is displayed by selecting **Show Info About Points** from the toolbar then holding the left-hand mouse button over the point shown. The orange color of this point indicates that the p-value for testing that there is no linkage disequilibrium between M9 and M14 is between 0.01 and 0.05. The pop-up window provides the exact value of this p-value.

Other parts of the plot can be viewed by selecting **Scroll Data** from the toolbar. Alternatively, the entire plot can be viewed in the window by selecting **Format →Rescale Axes** from the menu bar. This creates the following view of the plot:

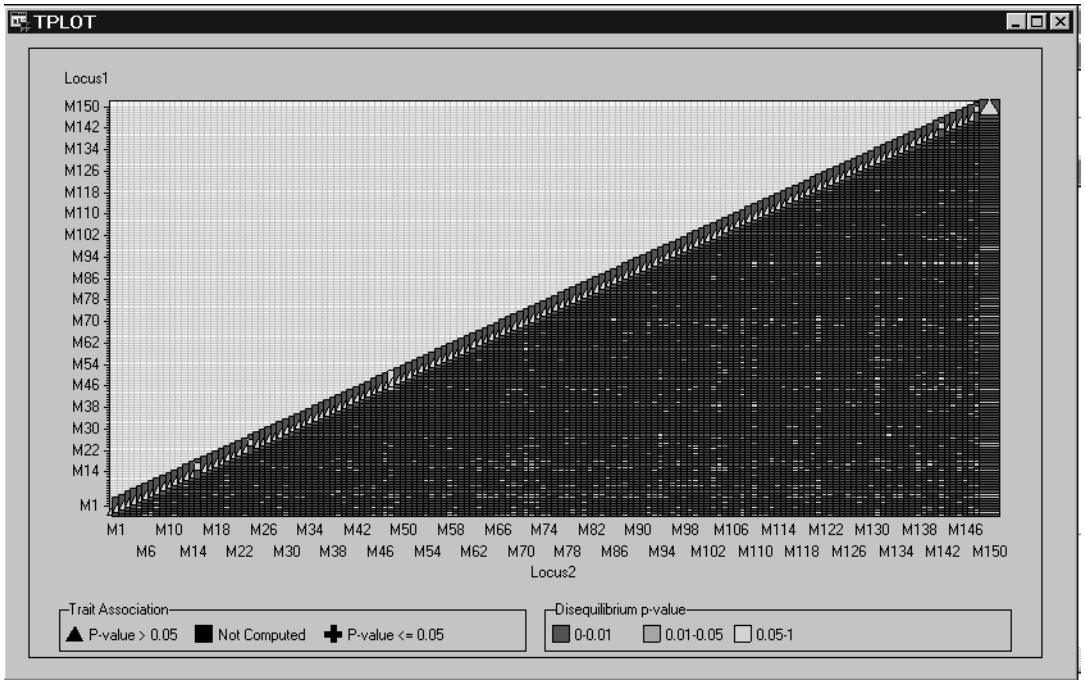


Figure 7.4. Results Window for TPLOT Macro

The view shown in Figure 7.4 displays all the data points at once.

Subject Index

Syntax Index

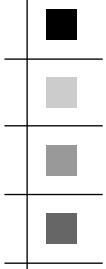
R

S

T

V

Y



www.sas.com/service/techsup/intro.html

SAS is the world leader in e-intelligence software and services. SAS partners with customers to turn raw data, including the vast quantity generated by e-business, into usable knowledge, and makes it available to decision-makers across the enterprise. SAS enables informed business decisions, helping its customers gain competitive advantage in their markets.