

An Introduction to Genetic Data Analysis Using SAS/Genetics™

Wendy Czika, Xiang Yu, and Russell D. Wolfinger
SAS Institute Inc.
Cary, North Carolina, USA

Abstract

The Human Genome Project and other recent genome research efforts have intensified the already vital relationship between the fields of genetics and statistics. One important application of the analysis of genetic data is locating genes that affect complex traits, human diseases in particular. While statistical procedures already in SAS® are capable of certain analyses, the unique nature of these data demands customized tools to fully accommodate the genetic dissection of a trait or disease. SAS/Genetics offers new statistical procedures specifically tailored for genetic data stored in a commonly used format. This paper discusses these methods and illustrates the features of the software with an example.

Introduction

Genetic markers, while often nonfunctional, play a key role in the mapping of human disease genes. Single nucleotide polymorphisms (SNPs) in particular are facilitating the creation of high-resolution genetic maps when used in linkage and association studies. SAS/Genetics provides a framework and the tools for analyzing markers and their relationship with a dichotomous trait (affection status of a disease e.g.). While many of these methods exploit the biallelic nature of SNPs in their design, multiallelic adaptations are available when studying markers such as microsatellites with more than two alleles.

Before examining marker-disease associations, initial calculations characterizing features of the markers should be performed. Properties of interest about each marker include measures of marker informativeness, which are useful for determining which markers may be most valuable for association or linkage studies; estimates of allele, genotype, and haplotype frequencies; tests for Hardy-Weinberg equilibrium (HWE); and measures of linkage disequilibrium (LD) between pairs of markers. These types of analyses can be accomplished with the ALLELE and HAPLOTYPE procedures. Once these marker features have been examined, marker-disease association tests can be performed on either random samples from populations of unrelated individuals who are either affected with a disease or unaffected, or on families with an affected child meeting certain requirements. Several different case-control and family-based tests can be implemented by the CASECONTROL and FAMILY procedures, respectively. Smoothing and multiple testing adjustments are then available to be applied to the p -values created by these tests in order to improve power and/or control type I error using the PSMOOTH procedure.

librium (HWE); and measures of linkage disequilibrium (LD) between pairs of markers. These types of analyses can be accomplished with the ALLELE and HAPLOTYPE procedures. Once these marker features have been examined, marker-disease association tests can be performed on either random samples from populations of unrelated individuals who are either affected with a disease or unaffected, or on families with an affected child meeting certain requirements. Several different case-control and family-based tests can be implemented by the CASECONTROL and FAMILY procedures, respectively. Smoothing and multiple testing adjustments are then available to be applied to the p -values created by these tests in order to improve power and/or control type I error using the PSMOOTH procedure.

Theory

Examining Marker Properties

The methods described in this section are discussed in detail by Weir (1996) unless otherwise cited. Some of the most basic descriptive statistics for genetic markers are the sample allele and genotype frequencies. A marker locus \mathbf{M} may have a series of alleles M_u , $u = 1, \dots, k$, which produce genotypes of the form M_u/M_v . In a sample of n individuals, the count of individuals with genotype M_u/M_v or M_v/M_u is n_{uv} . The number n_u of copies of allele M_u can then be found directly by summation: $n_u = 2n_{uu} + \sum_{v>u} n_{uv}$. The sample allele and genotype frequencies are written respectively as $\tilde{p}_u = n_u/(2n)$ and $\tilde{P}_{uv} = n_{uv}/n$. The variance of the sample allele frequency \tilde{p}_u is calculated using the multinomial distribution of genotype counts.

The variance of the sample genotype frequency \tilde{P}_{uv} is not generally calculated; instead, a maximum likelihood estimate (MLE) of the Hardy-Weinberg disequilibrium (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 its variance can be estimated using Fisher's approximate variance formula. Under ideal population conditions, the two alleles an individual receives, one from each parent, are independent so that $D_{uv} = 0$. 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 available for testing a marker for HWE, both which can accommodate any number of alleles. These methods are testing the hypothesis that $P_{uu} = p_u^2$ and $P_{uv} = 2p_u p_v$, $u \neq v$, that is $D_{uv} = 0$, for all $u, v = 1, \dots, k$.

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

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

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

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 $h = \sum_u \sum_{v>u} n_{uv}$ is the number of heterozygous individuals. Significance levels are calculated by the permutation procedure. The $2n$ alleles are randomly permuted to form new sets of n genotypes, and the proportion of times the value of the statistic from the permuted data exceeds the value from the actual data is the estimated significance level.

Three measures of marker informativeness are the polymorphism information content (PIC) (Botstein et al. 1980), (observed) heterozygosity, and allelic diversity (expected heterozygosity). They are calculated as follows:

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

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

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

These measures give an indication of the amount of heterozygosity in the sample at each marker. Markers with higher values of these measures tend to be more informative and of more use in association and linkage studies because there is more allelic variation.

Often, tests and measures based on haplotypes can be more powerful than those using genotypes alone. The set of genetic material an individual receives from each parent contains an allele at every locus, and a set of alleles on one chromosome is called a haplotype. However, it is often the case that haplotypes are unknown. You may have the genotypic information on an individual at several markers, for example you know an individual has genotype A/a at locus 1 and B/b at locus 2. But you do not know which set of haplotypes that individual received from his or her parents: A-B and a-b, or A-b and a-B. Clearly, as more loci are examined this problem becomes more complicated. Haplotype frequencies then must be estimated when haplotypes are not observed. For pairs of biallelic markers, the MLE of a haplotype frequency can be obtained analytically by solving a cubic equation. For multilocus haplotypes or markers with more than two alleles, an analytic solution is not tractable. The expectation-maximization (EM) algorithm, an iterative process, can be used to arrive at MLEs of haplotype frequencies, assuming HWE (Excoffier and Slatkin 1995; Hawley and Kidd 1995; Long, Williams, and Urbanek 1995).

In order to perform the EM algorithm, haplotype frequencies must first be initialized to some starting values. These values can be generated randomly, uniformly assigned, or calculated in a number of different ways. For a sample of n individuals, suppose the i th individual has genotype G_i . Genotype G_i has probability P_i in the population, which is calculated in the expectation step (E-step) as

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

where f_j is the frequency in the population of the j th possible haplotype h_j and $f_j^{c_i}$ is the frequency of the haplotype $h_j^{c_i}$ that together with h_j constitutes the genotype G_i . Now the log likelihood can be evaluated as

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

For each iteration after the first, the change in log likelihood from the previous iteration(s) is measured against a convergence criterion to determine whether the process needs to continue, or the maximum likelihood estimates have been reached. If the process must continue, the maximization step (M-step) is used as follows to create the haplotype frequencies for the next iteration:

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

where m_{ij} is the number of times haplotype h_j occurs in genotype G_i . The log likelihood will increase with each iteration of the EM algorithm until the maximum is reached. This maximum may be a local one, not a global maximum, and thus the EM algorithm can be performed several times using different sets of starting values for the haplotype frequencies; the best log likelihood will likely be the global maximum.

Haplotype frequencies, either observed or estimated, can be used in a variety of ways to glean more information from your marker data. The primary application is for examining linkage disequilibrium between markers. Haplotype frequencies can be expressed as $p_{uv} = p_u p_v + D_{uv}$ for haplotype M_u-N_v at loci **M** and **N**. D_{uv} is the linkage, or gametic, disequilibrium (LD) coefficient. When $D_{uv} = 0$ (the haplotype frequency is the product of the individual allele frequencies), the markers are said to be in linkage equilibrium, that is they are transmitted independently. Though there are many other factors that may affect disequilibrium, there is a general expectation that the amount of linkage disequilibrium is inversely related to the distance between the two loci. $D_{uv} = 0$ may be an indication that two loci are far apart on a chromosome, or even located on two different chromosomes. Chi-square approximations and permutation exact tests are available for testing whether $D_{uv} = 0$. When examining LD across more than two marker loci, a likelihood ratio test can be used for testing whether any linkage disequilibrium exists among the loci.

There are also five commonly used linkage disequilibrium measures that can be calculated for each pair of alleles M_u and N_v , described by Devlin and Risch (1995): the correlation coefficient r , the population attributable risk δ , Lewontin's D' , the proportional difference d , and Yule's Q . 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 together to represent one allele. Thus for each allele M_u in turn, \tilde{p}_1 will be 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 for "not N_v ." All measures have the same numerator, an estimate of $D = p_{11}p_{22} - p_{12}p_{21}$, which is equivalent to the LD coefficient for biallelic markers. The five measures are

$$\begin{aligned} r &= D / (p_1 p_2 q_1 q_2)^{1/2} \\ D' &= D / D_{\max} \\ \delta &= D / (q_1 p_{22}) \\ d &= D / (q_1 q_2) \\ Q &= D / (p_{11} p_{22} + p_{12} p_{21}) \end{aligned}$$

where $D_{\max} = \min(p_1 q_2, q_1 p_2)$ for $D > 0$ and $D_{\max} = \min(p_1 q_1, q_2 p_2)$ otherwise. Estimates of measures are calculated by replacing parameters with their appropriate estimates.

Exploring Marker-Trait Relationships

There are several different case-control tests that can be used to test for association between a genetic marker and a binary trait. With cases (affected individuals) and controls (unaffected individuals) representing the two rows, contingency tables can be created using genotype or allele categories for the columns. Tests for differences in these frequencies between case and control groups can be performed using the usual Pearson chi-square. Another type of test that can be performed is Armitage's trend test (1955), which uses the genotype contingency table in a different manner than Pearson's test; this is a test specifically for additive allele effects, whereas the Pearson chi-square based on the genotype contingency table is testing for both additive and dominance effects.

Table 1. Genotype Distribution for Case-Control Sample

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

Using [Table 1](#), the following chi-square statistics can be created:

$$\begin{aligned} X_T^2 &= \frac{N[N(r_1 + 2r_2) - R(n_1 + 2n_2)]^2}{R(N - R)[N(n_1 + 2n_2) - (n_1 + 2n_2)^2]} \\ X_A^2 &= \frac{2N[N(r_1 + 2r_2) - R(n_1 + 2n_2)]^2}{R(N - R)[2N(n_1 + 2n_2) - (n_1 + 2n_2)^2]} \\ X_G^2 &= \sum_{i=0}^2 \left[\frac{(Nr_i - Rn_i)^2}{NRn_i} + \frac{(Ns_i - Sn_i)^2}{NSn_i} \right] \end{aligned}$$

for the trend, allele, and genotype tests, which have 1, 1, and 2 df, respectively (Sasieni 1997). Extensions for multiallelic markers are straightforward for the allele and genotype tests: the Pearson chi-square test is carried out on the contingency table containing one column for each allele or genotype category. The degrees of freedom for a $2 \times m$ table is $(m - 1)$. Slager and Schaid (2001) offer a multiallelic version of the trend test to accommodate markers with more than two alleles; this test still has the same degrees of freedom as the allele test.

Contingency tables can also be created using haplotypes as columns. Since the number of haplotypes, and thus degrees of freedom, can grow quite large for multilocus haplotypes, an alternative approach is to construct for each haplotype one contingency table containing a column for that haplotype and a single column containing all other haplotypes. Directly observed or estimated haplotype counts can be used to calculate the Pearson statistic that has an asymptotic χ^2_1 distribution. Additionally, a likelihood ratio test based on the log likelihoods from the EM algorithm can be performed over all haplotypes at the markers of interest simultaneously. Such tests can be more powerful than the single-marker association tests.

The disadvantage with using case-control tests is that differences in allele, genotype, or haplotype frequencies between groups of cases and controls due to causes other than linkage may be found to be significant. That is, an association between the trait and marker may not be an indication of the marker's proximity to the trait gene. Family-based tests offer a solution to this problem: they test for linkage in the presence of association. These tests are performed using either family trios (parents and at least one affected child) or sib pairs consisting of one affected and one unaffected sibling.

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 , but are not discussed here.

The TDT (Spielman, McGinnis, and Ewens 1993) analyzes families where both parents have been genotyped for the marker and at least one is heterozygous. 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 bino-

mial 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.

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 is 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_1M_1 and s that are M_1M_2 , summing over families gives

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

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

as the expected value and variance of Y , respectively.

The S-TDT and TDT can be combined for situations when some, but not all, of the parents have been genotyped: the TDT is applied to all families that meet its requirements. The S-TDT is then applied to the remaining families that meet the 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

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

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, define for each family in the data the average number of M_1 alleles among affected siblings and unaffected siblings respectively as $m^a = Y/a$ and $m^u = [(2r + s) - Y]/u$. Then for each family,

$d = m^a - m^u$ and summing over each family gives $S = \sum \text{sgn}(d)$, where $\text{sgn}(d)=1$ for $d > 0$, 0 for $d = 0$, and -1 for $d < 0$. The SDT statistic is then defined as $T = S^2/W$ where $W = \sum (\text{sgn}(d))^2$. This statistic has an asymptotic χ_1^2 distribution, and this is used to obtain p -values for the SDT. This sibship test can also be combined with the TDT, creating a test that can potentially use more of the data (Horvath and Laird 1998).

The RC-TDT (Knapp 1999) takes the combined S-TDT a step further by reconstructing missing parental genotypes when possible in order to make use of 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 take into account the bias created by the genotype reconstruction, to form the z score over all families $z = (Y - e)/\sqrt{v}$.

Adjusting p -Values

In association studies, there can potentially be thousands of markers being analyzed simultaneously. In addition to the multiple testing issues that must be taken into consideration, there is also the fact that many of these tests are correlated due to LD. This correlation can be used as a smoothing tool to help eliminate false-positives that are found. Various methods can be used to combine p -values from neighboring markers into a new p -value for each marker by implementing the sliding window approach described by Zaykin et al. (2002). Since this method does not change the number of tests being performed, multiple testing corrections such as Bonferroni and Sidak (1967) can then be applied to the new set of p -values.

Data Analysis

To help demonstrate the statistical methods that are available in the SAS/Genetics software, a sample

analysis of simulated data from GAW12 (Wijsman et al. 2001) is presented. To perform the analysis, you need to create data sets from three separate data files. The first contains the marker genotype data; this file is named Genmrk19.1 and contains individuals from a replicate of the general population genotyped at 105 markers from chromosome 19. Here are the first 10 rows and several columns of the data:

```

1  1  4/ 7  1/ 5  7/ 1  1/ 1  2/ 3 ...
1  2  8/ 7  2/ 7  4/ 7  4/ 8  3/ 3 ...
1  3  9/ 2  5/ 1  6/ 4  3/ 5  3/ 3 ...
1  4  2/ 3  1/ 6  3/ 3  1/ 3  4/ 4 ...
1  5  3/ 4  7/ 3  7/ 6  5/ 3  1/ 2 ...
1  6  7/ 8  4/ 1  7/ 3  8/ 3  4/ 4 ...
1  7  3/ 9  5/ 4  6/ 3  5/ 5  3/ 2 ...
1  8  1/ 8  5/ 1  8/ 8  4/ 3  3/ 4 ...
1  9  3/ 3  5/ 1  7/ 7  3/ 9  3/ 3 ...
1 10  8/ 4  4/ 4  2/ 7  3/ 2  1/ 4 ...

```

The first two columns contain the pedigree and individual IDs, respectively. They are followed by genotypes at each of 105 markers (only 5 shown here), with the two alleles comprising each marker genotype separated by a "/". The actual data set contains 1497 rows, each representing an individual. These data can easily be read into a SAS data set using the following code:

```

data markers;
  infile 'Genmrk19.1' delimiter='/' lrecl=640;
  input ped id a1-a210;
run;

```

This data set needs to be combined with the phenotype information, contained in the file Genpheno.1. Here are the first 15 rows of this file:

```

1  1  0  0  1  D
1  2  0  0  2  D
1  3  0  0  2  D
1  4  0  0  1  D
1  5  0  0  1  D
1  6  0  0  2  D
1  7  0  0  2  D
1  8  0  0  2  D
1  9  0  0  1  D
1 10  0  0  2  A 37 320  37.78 0  U  16.15 ...
1 11  0  0  1  A 40  2 269.39 0  U  10.41 ...
1 12  0  0  1  A 40  3  15.28 0  A 40 12.86 ...
1 13  0  0  1  A 49  4  14.16 0  U  15.63 ...
1 14  0  0  1  A 64  5  33.17 1  U  17.52 ...
1 15  0  0  1  A 65  6  45.96 1  U  21.63 ...

```

The first two columns of these data are the same as those in the marker data, the ID variables. They are followed by two columns of parental IDs identifying the father and mother of the individual. The next column represents gender, and the sixth column contains a "D" for deceased individuals and "A" for alive individuals. Note that more columns follow only for living individuals. Of these columns, only the fifth column following the living status column is needed in this analysis; this column represents the disease status of the individual, "A" for individuals affected with the disease

and “U” for unaffected individuals. These data can be read into a SAS data set and then combined with the marker data as follows:

```
data pheno;
  infile 'Genpheno.1' missover;
  input ped id father mother gender living $ age
        hh ef1 ef2 status $ age_o 45-46 q1-q5;
  keep ped id father mother status;
run;

proc sort data=markers;
  by ped id;
run;

data markertrt;
  merge pheno markers;
  by ped id;
run;
```

Additionally, you can use the marker names provided. Here are the first 10 of 105 rows of the file Map19:

```
D19G001 0.70
D19G002 1.42
D19G003 2.51
D19G004 4.50
D19G005 4.87
D19G006 6.24
D19G007 6.46
D19G008 7.09
D19G009 10.85
D19G010 11.51
```

which can be read into a SAS data set using the code

```
data map;
  infile 'Map19';
  input name $ location;
run;
```

This data set can be specified in the NDATA= option of the ALLELE, HAPLOTYPE, CASECONTROL, and FAMILY procedures to provide names for the markers in the displayed output and output data sets these procedures create. These four procedures require a VAR statement that identifies the columns of marker alleles to analyze, with the two columns of alleles for each marker listed consecutively.

For the ALLELE, HAPLOTYPE, and CASECONTROL procedures, which assume individuals have been randomly sampled from the population, a subset of the data containing only founders (identified by having the value 0 for each of their parental IDs) who are living are used.

Examining Marker Properties

The ALLELE procedure shown in the following code calculates some basic summary statistics for each marker included in the analysis.

```
proc allele data=founders ndata=map outstat=ld nofreq
  exact=10000 haplo=est corrcoeff dprime;
  var a1-a210;
```

```
run;

proc print data=ld noobs;
run;
```

Since there are 105 markers in this data set, the VAR statement contains 210 variables. By default, three tables will be created. The first is a marker summary table, containing measures of marker informativeness: the polymorphism information content (PIC), heterozygosity, and allelic diversity; the number of alleles and number of individuals typed at each marker; and statistics for the HWE test. The test for HWE is performed by default using the asymptotic chi-square test; in this analysis, the exact version is performed as well by permuting the marker alleles, with the number of permutations indicated by the EXACT= option. Here is the “Marker Summary” table, shown for the first 5 markers only:

The ALLELE Procedure					
Marker Summary					
Locus	Number of Indiv	Number of Alleles	PIC	Heterozygosity	Allelic Diversity
D19G001	165	9	0.8235	0.8606	0.8430
D19G002	165	7	0.7873	0.8061	0.8137
D19G003	165	8	0.8011	0.8242	0.8233
D19G004	165	9	0.7986	0.8242	0.8211
D19G005	165	4	0.5956	0.7273	0.6560

Marker Summary					
-----Test for HWE-----					
Locus	Chi-Square	DF	Pr > ChiSq	Prob Exact	
D19G001	21.3950	36	0.9744	0.9464	
D19G002	27.4028	21	0.1579	0.1941	
D19G003	23.1209	28	0.7270	0.8453	
D19G004	23.5607	36	0.9451	0.9525	
D19G005	7.7220	6	0.2592	0.3280	

The next two tables are the allele and genotype frequency tables. In addition to the frequency estimates themselves, these tables contain an estimate of the standard deviation of the frequencies. When the BOOTSTRAP= option is included in the PROC ALLELE statement, bootstrap confidence intervals for the frequency estimates are formed using the number of bootstrap samples specified in the option and a confidence level of 0.05 by default, or the level that is provided in the ALPHA= option. These tables have been suppressed in this analysis with the NOFREQ option.

An output data set is created with the OUTSTAT= option in the PROC ALLELE statement. This data set contains the statistics from the HWE tests in the rows where Locus1 and Locus2 are the same, and statistics for a test for LD between pairs of markers in the rows where these two variables differ. Since the HAPLO=EST option is included in this code, the LD test and measures are calculated using maximum likelihood estimates of the two-locus haplotype frequencies for each pair of alleles at the two loci. Only

the asymptotic chi-square test is used because there is no exact test available when haplotype frequencies are estimated. The following SAS output shows a subset of this output data set:

OUTSTAT= Data Set from PROC ALLELE						
Locus1	Locus2	NIndiv	Test	ChiSq	DF	ProbChi
D19G001	D19G001	165	HWE	21.3950	36	0.97441
D19G001	D19G002	165	LD	84.4138	48	0.00091
D19G001	D19G003	165	LD	68.0223	56	0.13017
D19G001	D19G004	165	LD	60.9314	64	0.58569
D19G001	D19G005	165	LD	34.6481	24	0.07379
D19G002	D19G002	165	HWE	27.4028	21	0.15791
D19G002	D19G003	165	LD	49.8788	42	0.18866
D19G002	D19G004	165	LD	54.4225	48	0.24330
D19G002	D19G005	165	LD	24.1505	18	0.15016
D19G003	D19G003	165	HWE	23.1209	28	0.72696

The five LD measures previously described are offered by PROC ALLELE and calculated for each pair of alleles at two loci when requested. The table reporting these measures is only included in the output when one or more of these measures have been specified. Below is a subset of the "Linkage Disequilibrium Measures" table that displays the estimated haplotype frequencies, LD coefficients, and the two requested measures, the correlation coefficient r and Lewontin's D' .

The ALLELE Procedure							
Linkage Disequilibrium Measures							
Locus1	Locus2	Haplotype	Frequency	LD Coeff	Corr Coeff	Lewontin's D'	
D19G071	D19G072	1-1	0.0360	0.008	0.064	0.205	
D19G071	D19G072	1-2	0.1198	-0.016	-0.067	-0.115	
D19G071	D19G072	1-3	0.0968	0.020	0.106	0.192	
D19G071	D19G072	1-4	0.1711	-0.010	-0.041	-0.056	
D19G071	D19G072	2-1	0.0110	-0.009	-0.080	-0.454	
D19G071	D19G072	2-2	0.1323	0.035	0.163	0.170	
D19G071	D19G072	2-3	0.0575	0.002	0.013	0.019	
D19G071	D19G072	2-4	0.0940	-0.036	-0.160	-0.279	
D19G071	D19G072	3-1	0.0000	-0.006	-0.086	-1.000	
D19G071	D19G072	3-2	0.0195	-0.011	-0.078	-0.353	
D19G071	D19G072	3-3	0.0196	0.003	0.023	0.033	
D19G071	D19G072	3-4	0.0528	0.012	0.086	0.231	
D19G071	D19G072	4-1	0.0165	0.004	0.045	0.080	
D19G071	D19G072	4-2	0.0568	-0.002	-0.009	-0.027	
D19G071	D19G072	4-3	0.0117	-0.021	-0.143	-0.645	
D19G071	D19G072	4-4	0.1008	0.023	0.118	0.218	

In order to accommodate the frequency estimation of haplotypes spanning more than two loci, SAS/Genetics also offers the HAPLOTYPED procedure. This procedure implements the EM algorithm to calculate maximum likelihood estimates of the haplotype frequencies assuming HWE with the following code:

```
proc haplotype data=founders cutoff=0.01 out=outhap;
  var a1-a10;
  trait status;
run;
```

Here, haplotypes at the first five markers only are examined. One of the tables included in the output by default is the "Haplotype Frequencies" table. The results from using the TRAIT statement are shown

in the following section. The CUTOFF= option requests that only haplotypes with an estimated frequency greater than the number specified be displayed. As shown in the following table, out of more than 18,000 possible haplotypes, only 12 have an estimated frequency greater than 0.01:

The HAPLOTYPED Procedure					
Haplotype Frequencies					
Number	Haplotype	Freq	Standard Error	95% Confidence Limits	
1	2-1-7-3-4	0.01212	0.00603	0.00030	0.02395
2	2-1-7-8-4	0.01212	0.00603	0.00030	0.02395
3	5-5-7-9-3	0.01212	0.00603	0.00030	0.02395
4	7-4-6-4-3	0.01212	0.00603	0.00030	0.02395
5	8-1-8-3-4	0.01212	0.00603	0.00030	0.02395
6	8-3-3-9-4	0.01479	0.00665	0.00174	0.02783
7	8-3-7-3-3	0.01515	0.00673	0.00195	0.02835
8	8-5-3-3-3	0.01818	0.00737	0.00375	0.03262
9	9-4-6-3-3	0.01212	0.00603	0.00030	0.02395
10	9-5-4-3-3	0.02121	0.00794	0.00564	0.03678
11	9-5-7-6-3	0.01212	0.00603	0.00030	0.02395
12	9-6-4-3-4	0.01212	0.00603	0.00030	0.02394

The OUT= data set provides for each possible haplotype pair within an individual given the individual's genotype, the probability that the haplotype pair comprises that genotype. Here, the first 10 rows of this output data set are shown:

OUT= Data Set from PROC HAPLOTYPED													
ID	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	HAPLOTYPED1	HAPLOTYPED2	PROB
1	8	4	4	4	2	7	3	2	1	4	4-4-7-2-1	8-4-2-3-4	1.00000
2	5	9	3	5	3	4	2	3	4	3	5-3-3-2-4	9-5-4-3-3	1.00000
3	8	2	5	1	6	3	3	5	3	4	2-1-6-5-4	8-5-3-3-3	1.00000
4	7	8	5	3	8	4	5	3	3	4	7-5-8-3-3	8-3-4-5-4	1.00000
5	9	2	2	5	7	6	9	3	2	4	2-2-6-9-2	9-5-7-3-4	1.00000
6	2	7	1	4	6	7	8	4	4	3	2-1-7-8-4	7-4-6-4-3	1.00000
7	7	7	6	6	1	4	9	5	3	1	7-6-1-5-1	7-6-4-9-3	0.25000
7	7	7	6	6	1	4	9	5	3	1	7-6-1-5-3	7-6-4-9-1	0.25000
7	7	7	6	6	1	4	9	5	3	1	7-6-1-9-1	7-6-4-5-3	0.25000
7	7	7	6	6	1	4	9	5	3	1	7-6-1-9-3	7-6-4-5-1	0.25000

The first six individuals have only one possible haplotype pair given their genotype based on the inferred haplotype frequencies. For the seventh individual, there are four equally probable haplotype pairs.

Exploring Marker-Trait Relationships

The CASECONTROL, FAMILY, and HAPLOTYPED procedures are available for testing for associations between dichotomous traits and markers. The following code demonstrates how to implement PROC CASECONTROL:

```
proc casecontrol data=founders ndata=map outstat=cctestst
  genotype trend;
  var a1-a210;
  trait status;
run;

proc print data=cctestst (obs=10) noobs;
run;
```

In addition to the VAR statement previously described, PROC CASECONTROL has a TRAIT statement where a dichotomous trait such as disease

status is identified. Three options are available in the PROC CASECONTROL statement that indicate which case-control tests should be performed. The available tests are the allele case-control test (not specified here), the genotype case-control test, and the linear trend test. Note that the case-control statistics all test the hypothesis of no association between each marker and trait. The output data set that is created by this procedure contains the test statistics for each test and each marker. A subset of the output data set is displayed here:

OUTSTAT= Data Set from PROC CASECONTROL						
Locus	ChiSq Genotype	ChiSq Trend	df Genotype	df Trend	Prob Genotype	Prob Trend
D19G001	32.3041	5.8754	44	8	0.90414	0.66118
D19G002	22.9640	7.5170	27	6	0.68695	0.27566
D19G003	25.3857	5.0110	35	7	0.88367	0.65862
D19G004	26.7125	4.2236	44	8	0.98158	0.83641
D19G005	13.0445	0.8921	9	3	0.16060	0.82732
D19G006	74.4421	21.0026	119	14	0.99954	0.10157
D19G007	23.4462	9.4603	27	6	0.66082	0.14930
D19G008	31.7414	6.8149	35	7	0.62621	0.44841
D19G009	7.0928	3.3052	9	3	0.62746	0.34692
D19G010	47.1247	9.0602	77	11	0.99712	0.61634

The FAMILY procedure has a similar syntax to that of the CASECONTROL procedure, with the only differences being the tests and options offered in the PROC FAMILY statement, an option available in the TRAIT statement, and the addition of an ID statement to designate the names of the individual and family identifiers, as shown in the following code:

```
proc family data=markertrt ndata=map outstat=famtests
    tdt contcorr;
    var a1-a210;
    trait status / affected='A';
    id ped id father mother;
run;

proc print data=famtests (obs=10) noobs;
run;
```

For the family-based tests, the value of the TRAIT variable that is considered “affected” can have a significant impact on the test results. By default, PROC FAMILY uses the second value of the TRAIT variable encountered in the input data set as the value corresponding to affected; alternatively, to ensure that the proper value is used, you can specify the appropriate value in the /AFFECTED= option of the TRAIT statement. As seen in the code, up to four identifiers may be included in the ID statement: the pedigree, individual, and two parental IDs. The pedigree variable need not be specified if all individual IDs are unique. There are several family-based tests available including those to accommodate missing parental genotypes, but only the TDT is used in this analysis since the parental genotypes have been obtained. An output data set with a similar structure to the one created by PROC CASECONTROL is produced by PROC FAMILY.

OUTSTAT= Data Set from PROC FAMILY			
Locus	ChiSq TDT	df TDT	ProbTDT
D19G001	7.9903	8	0.43441
D19G002	8.2224	6	0.22226
D19G003	6.4880	7	0.48406
D19G004	1.7472	8	0.98780
D19G005	1.8750	3	0.59875
D19G006	12.7737	14	0.54442
D19G007	4.1984	6	0.64985
D19G008	9.1880	7	0.23944
D19G009	2.8388	3	0.41715
D19G010	7.4416	11	0.76228

In the previous section, a TRAIT statement was specified in PROC HAPLOTYPE like the one in the CASECONTROL and FAMILY procedures. This invokes the likelihood ratio test for marker-trait association using the estimated haplotype frequencies, performed over all haplotypes at the markers. The results are displayed in the “Test for Marker-Trait Association” table:

Test for Marker-Trait Association						
Trait Number	Trait Value	Num Obs	DF	LogLike	Chi- Square	Pr > ChiSq
1	U	120	1763	-1132		
2	A	45	812	-360.27066		
	Combined	165	2281	-1620	127.7227	1.0000

This table shows the p -value from the χ^2_{294} distribution. An exact p -value can also be calculated by creating new samples with the trait values permuted and comparing the chi-square statistic of the new samples to the one from the original sample.

Adjusting p -Values

In this example, the case-control and family-based tests are each performed on over 100 markers. The PSMOOTH procedure offers two types of p -value adjustments for these test results: smoothing methods to take into account p -values from neighboring, and likely correlated, markers and multiple testing corrections to account for the number of tests being performed. The following code invokes Simes’ method (1986) for smoothing the p -values using a bandwidth of 10, which translates into a sliding window containing 21 markers. The two methods available to correct for multiple testing are Bonferroni and Sidak. These methods, had they been included as options in the PROC statement, would have adjusted the smoothed p -values according to the number of markers tested (105 in this scenario). The VAR statement indicates which column of p -values to adjust, and the ID statement lists the variables to include in the output data set along with the original and adjusted p -values.

```
proc psmooth data=famtests out=sm_famtests simes bw=10;
    var ProbTDT;
    id Locus;
run;

proc print data=sm_famtests (obs=10) noobs;
run;
```

The following is a subset of the output data set created by PROC PSMOOTH:

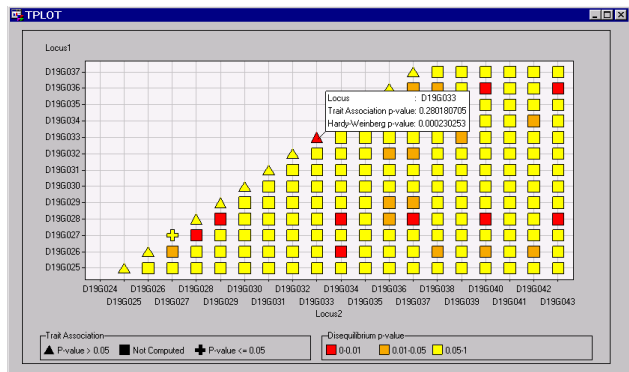
OUT= Data Set from PROC PSMOOTH		
Locus	ProbTDT	ProbTDT_ S10
D19G001	0.43441	0.79427
D19G002	0.22226	0.86647
D19G003	0.48406	0.90088
D19G004	0.98780	0.88933
D19G005	0.59875	0.95285
D19G006	0.54442	0.93819
D19G007	0.64985	0.92063
D19G008	0.23944	0.95115
D19G009	0.41715	0.94493
D19G010	0.76228	0.92263

Evaluating Results

Instead of forcing you to scrutinize four separate output data sets, SAS/Genetics includes the %TPLOTT macro, which creates a graphical representation of the genetic marker test results. This plot contains different symbols and colors of the symbols representing certain p -value ranges. On a single plot, results from testing single markers for HWE and associations with the trait and pairs of markers for LD are combined. Here is a sample invocation of the macro using the output data sets that have been created in the preceding examples:

```
%tplot( ld, famtests, ProbTDT );
```

The first argument to the %TPLOTT macro is the output data set from PROC ALLELE that contains the HWE and LD test results. The second argument is the output data set from PROC FAMILY containing p -values for the family-based association test, the TDT. Note that output data sets from either PROC CASECONTROL or PROC PSMOOTH can also be used for the second argument, or any user-created data set containing p -values. The third argument indicates which column of p -values from the data set specified in the second argument to use in the plot. The following is the color plot that is produced by this macro:



The color of the squares on the off-diagonal of the plot indicates the p -value range for the LD test. The symbols on the diagonal represent both the HWE p -value range using color, and the marker-disease association p -value range from the TDT using the shape of the symbol. In order to see the exact values of the p -values, the pop-up window can be viewed as displayed. This plot provides, at a glance, an indication of the extent that LD between markers extends, which regions show a significant association with disease status, and for which markers Hardy-Weinberg proportions hold.

Conclusion

As shown in the preceding example, SAS/Genetics offers resources for analyzing genetic marker data that go well beyond the methods currently available in SAS. Procedures specifically designed for marker data provide a seamless transition from data collection to analysis and supply tools needed for the association mapping of a complex trait or disease. Gene-mapping is further facilitated by a graphical representation of p -values in a distinct triangular plot.

References

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

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.

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.

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

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

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.

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.

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.

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

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.

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.

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.

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

Wijsman, E.M., et al. (2001), “Analysis of Complex Genetic Traits: Applications to Asthma and Simulated Data,” *Genetic Epidemiology*, 21, S1–S853.

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.

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

Acknowledgments

The simulated GAW12 data are supported by NIGMS grant GM31575.

Contact Information

Wendy Czika
SAS Institute Inc.
R-3212 SAS Campus Drive
Cary, NC 27513

Email: wendy.czika@sas.com