The procedures in SAS/Genetics (Release 9.1.3)

Presented by Tõnu Möls
Journal Club: 24th March 2009

8 procedures:

Procedure	**Implementation**
ALLELE	Wendy Czika
CASECONTROL	Wendy Czika
FAMILY	Wendy Czika
HAPLOTYPE	Xiang Yu
HTSNP	Xiang Yu
INBREED	Anthony Baiching An, Meltem Narter
PSMOOTH	Wendy Czika
TPLOT Macro	Wendy Czika
TPLOT Results (Frame)	Art Barnes
TPLOT Results (SCL)	Susan E. Haller

ALLELE procedure

Preliminary analyses on genetic marker data.

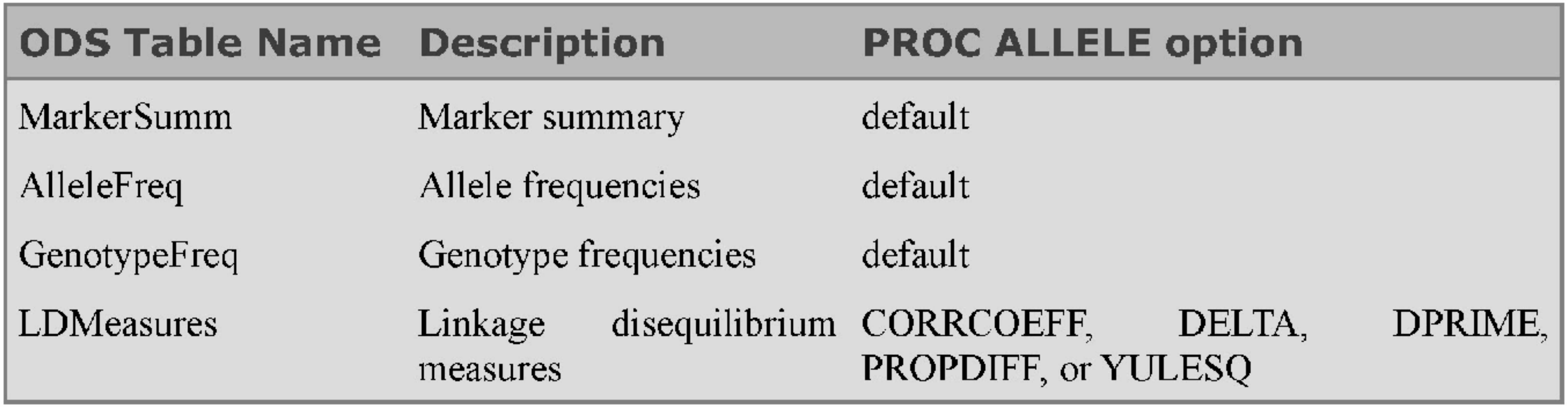
Joint analyses on markers and traits

Multinomial distribution of marker alleles

Random sampling

Indication of marker informativeness

ODS Tables Created by the ALLELE Procedure



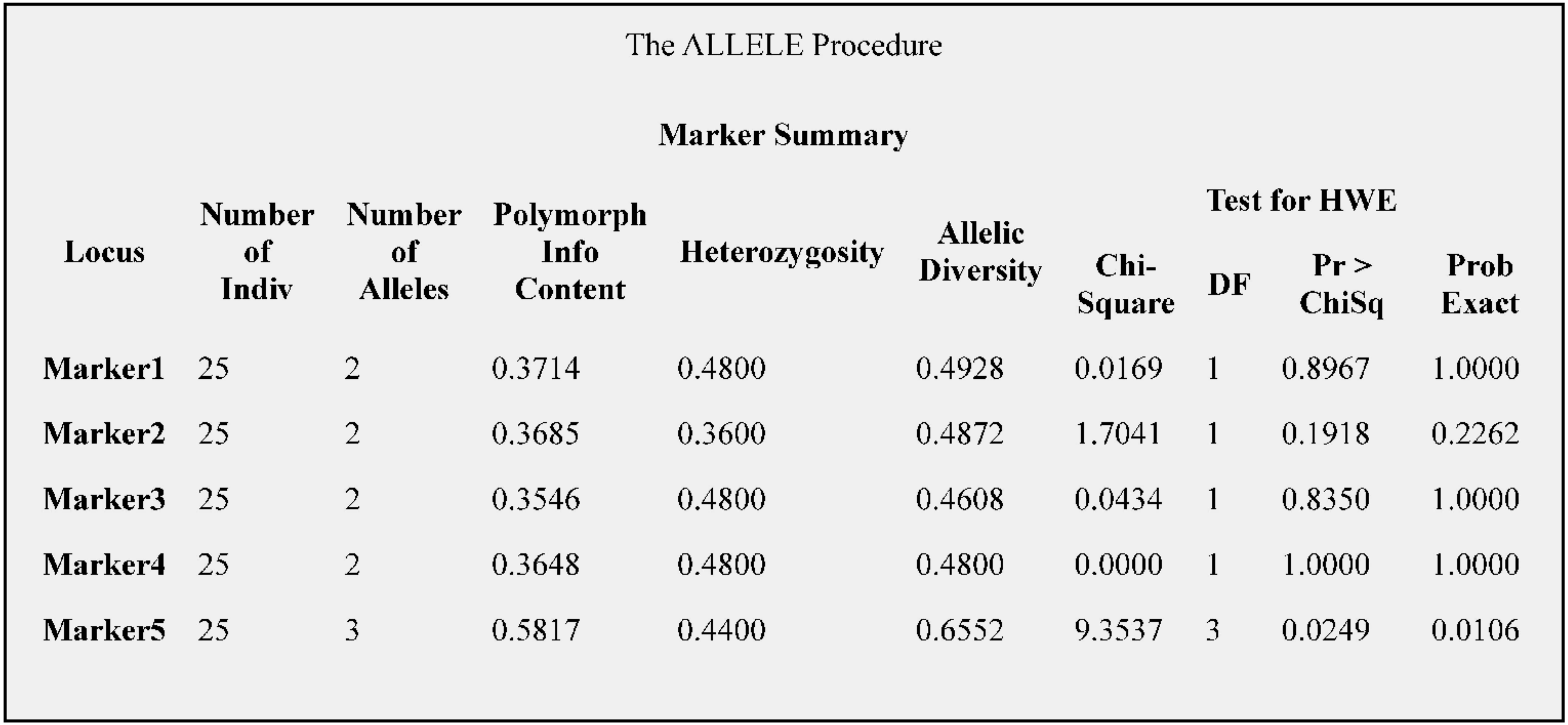
ODS Table Name	Description	PROC ALLELE option
MarkerSumm	Marker summary	default
AlleleFreq	Allele frequencies	default
GenotypeFreq	Genotype frequencies	default
LDMeasures	Linkage disequilibrium measures	CORRCOEFF, DELTA, DPRIME, PROPDIFF, or YULESQ

```
data markers;
      input (a1-a10) ($);
      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
   ;
```

```
data markers;
   input (g1-g5) ($);
   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
 ;
```

Marker Summary for the ALLELE Procedure



The ALLELE Procedure

Marker Summary

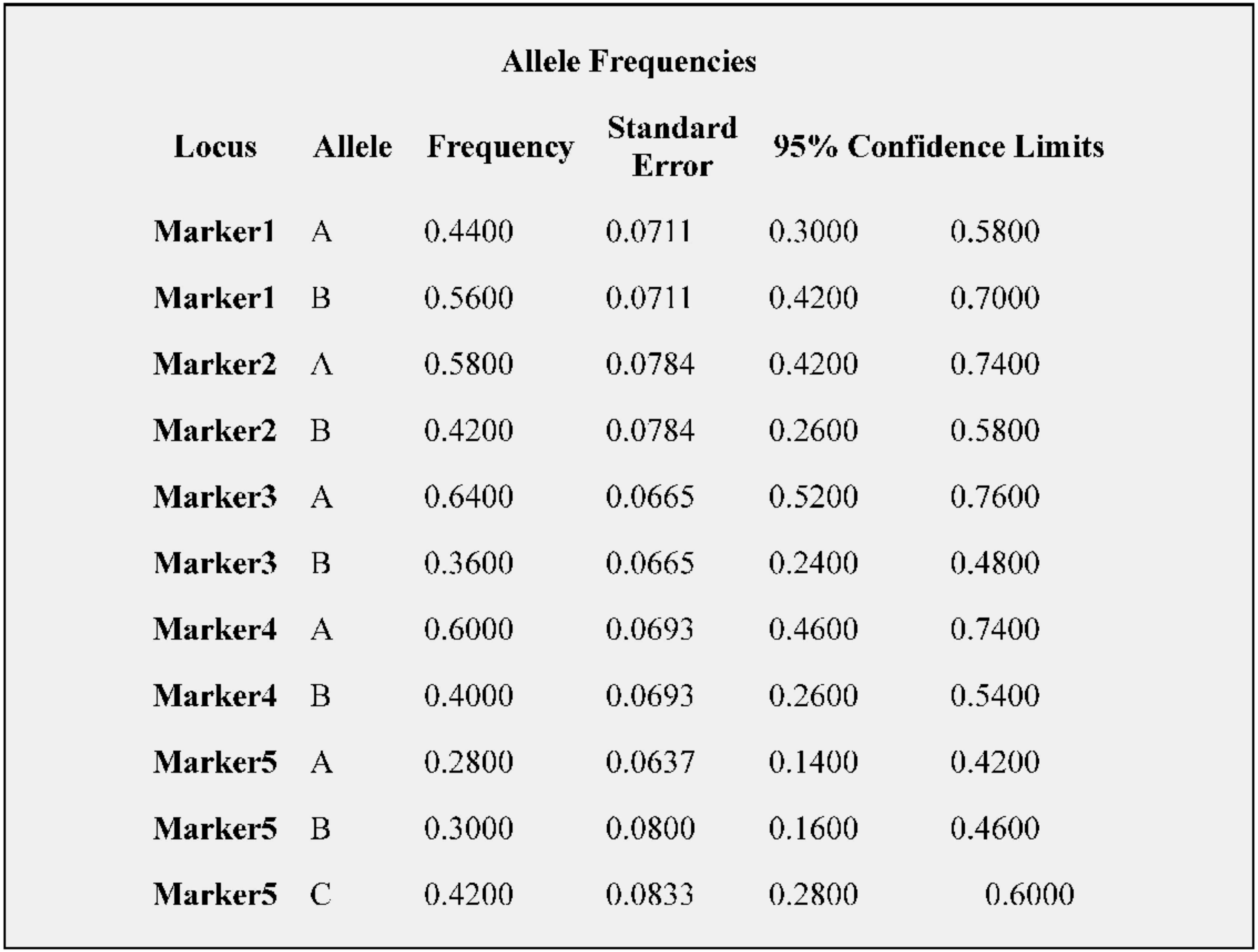
Locus	Number of Indiv	Number of Alleles	Polymorph Info Content	Heterozygosity	Allelic Diversity	Test for HWE			
						Chi-Square	DF	Pr > ChiSq	Prob Exact
Marker1	25	2	0.3714	0.4800	0.4928	0.0169	1	0.8967	1.0000
Marker2	25	2	0.3685	0.3600	0.4872	1.7041	1	0.1918	0.2262
Marker3	25	2	0.3546	0.4800	0.4608	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

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

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

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

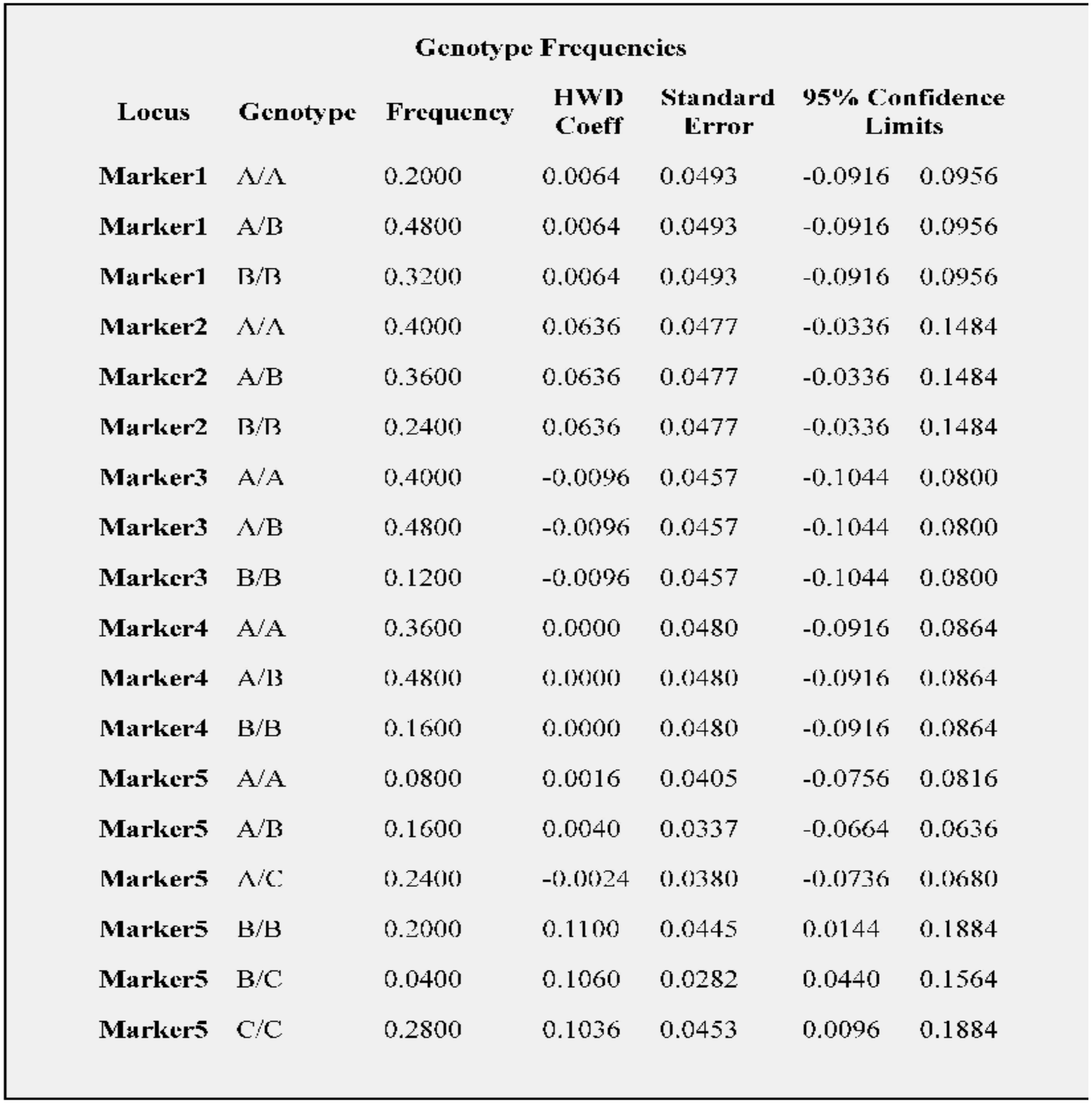
ODS table of allele frequencies



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

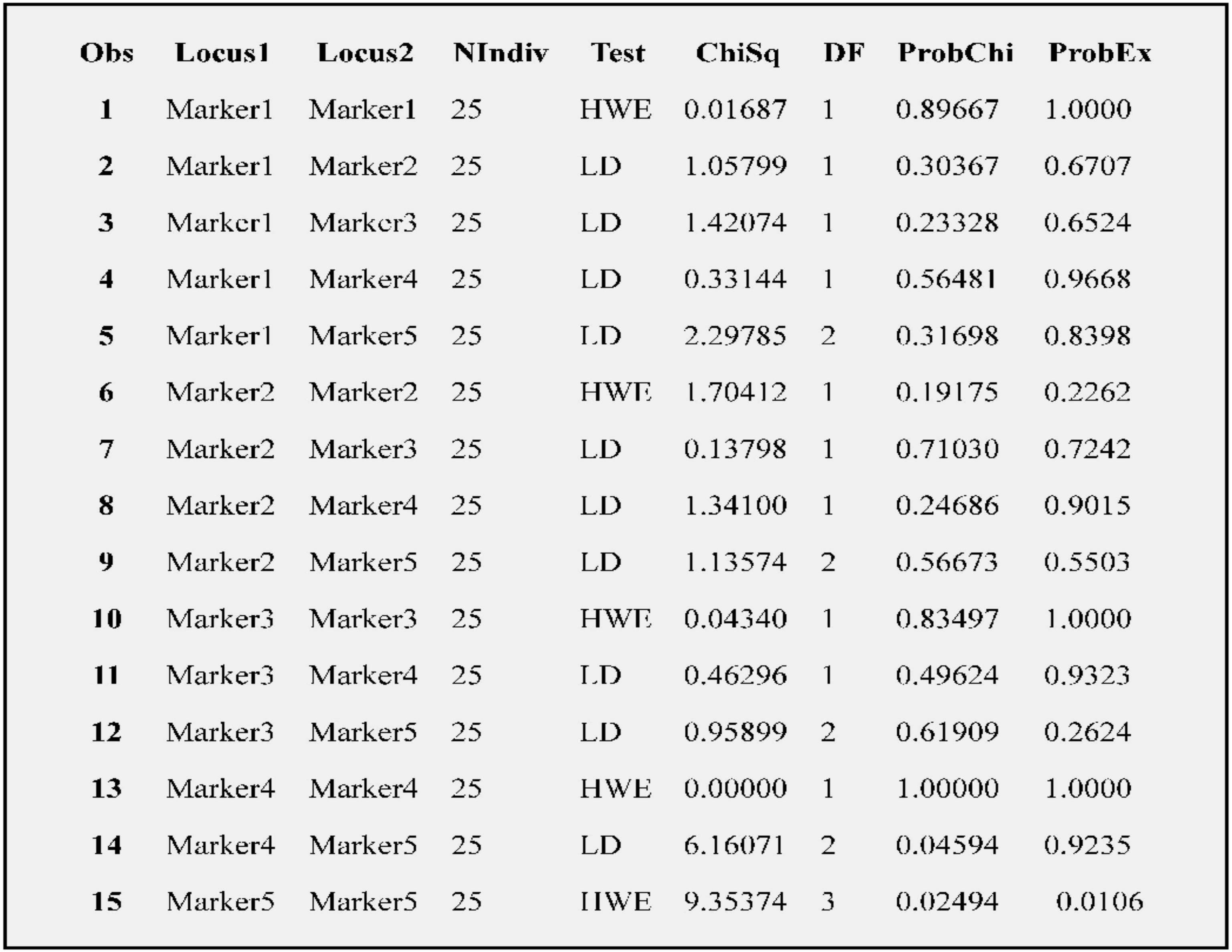
Genotype frequencies for each marker with the associated disequilibrium coefficient, its standard error, and the 95% confidence limits



Genotype Frequencies

Locus	Genotype	Frequency	HWD Coeff	Standard Error	95% Confidence Limits	
Marker1	A/A	0.2000	0.0064	0.0493	-0.0916	0.0956
Marker1	A/B	0.4800	0.0064	0.0493	-0.0916	0.0956
Marker1	B/B	0.3200	0.0064	0.0493	-0.0916	0.0956
Marker2	A/A	0.4000	0.0636	0.0477	-0.0336	0.1484
Marker2	A/B	0.3600	0.0636	0.0477	-0.0336	0.1484
Marker2	B/B	0.2400	0.0636	0.0477	-0.0336	0.1484
Marker3	A/A	0.4000	-0.0096	0.0457	-0.1044	0.0800
Marker3	A/B	0.4800	-0.0096	0.0457	-0.1044	0.0800
Marker3	B/B	0.1200	-0.0096	0.0457	-0.1044	0.0800
Marker4	A/A	0.3600	0.0000	0.0480	-0.0916	0.0864
Marker4	A/B	0.4800	0.0000	0.0480	-0.0916	0.0864
Marker4	B/B	0.1600	0.0000	0.0480	-0.0916	0.0864
Marker5	A/A	0.0800	0.0016	0.0405	-0.0756	0.0816
Marker5	A/B	0.1600	0.0040	0.0337	-0.0664	0.0636
Marker5	A/C	0.2400	-0.0024	0.0380	-0.0736	0.0680
Marker5	B/B	0.2000	0.1100	0.0445	0.0144	0.1884
Marker5	B/C	0.0400	0.1060	0.0282	0.0440	0.1564
Marker5	C/C	0.2800	0.1036	0.0453	0.0096	0.1884

Statistics for testing individual markers for HWE and marker pairs for linkage disequilibrium LD



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	IIWE	9.35374	3	0.02494	0.0106

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

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

Heterozygosity

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

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

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

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

Testing for Hardy-Weinberg Equilibrium

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

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

Permutation Version of Exact Test

The permutation version of 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

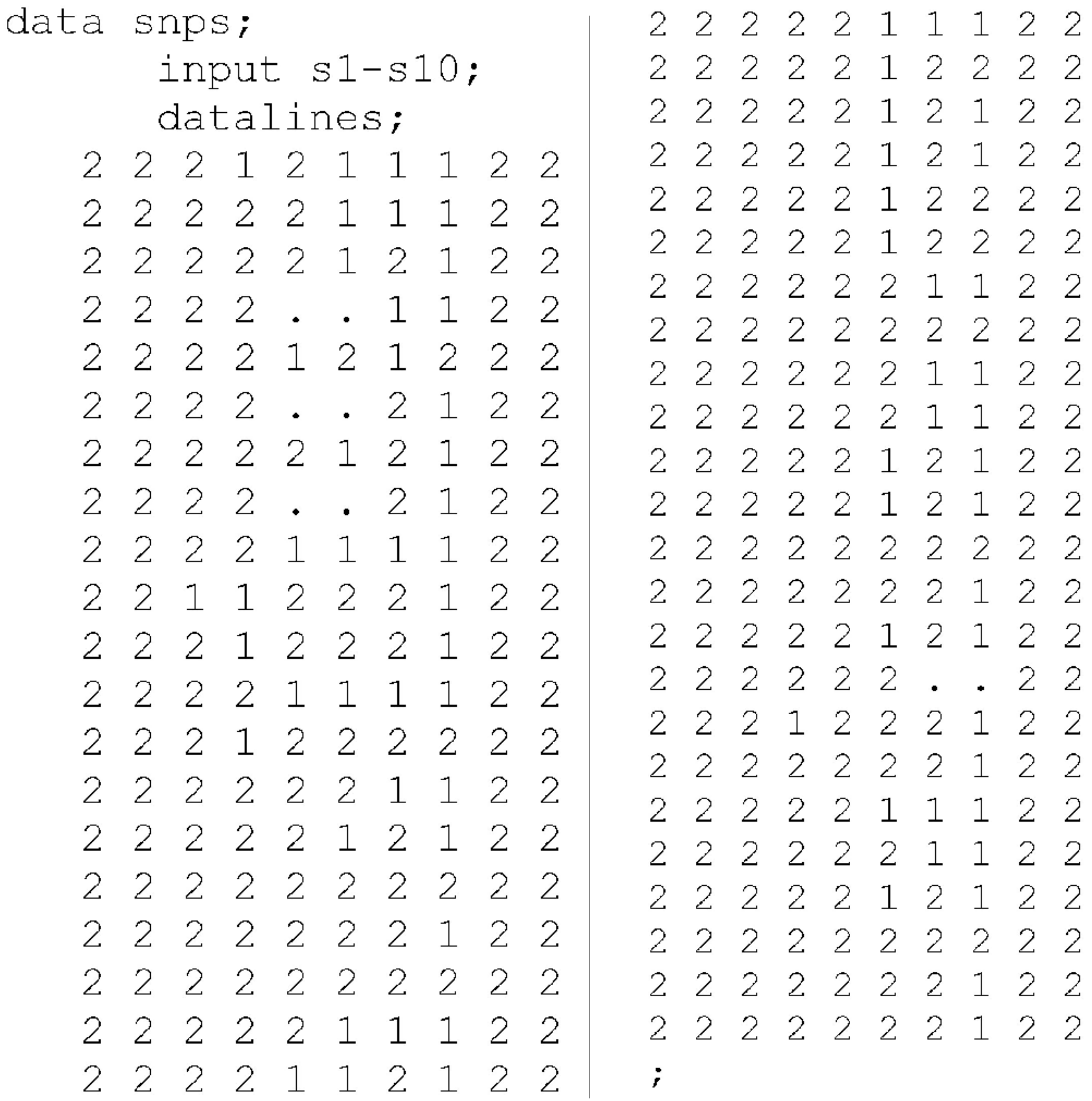
$$h = \sum_u \sum_{v \neq u} n_{uv}$$

is the number of heterozygous individuals. Significance levels are calculated by the Monte Carlo permutation procedure. The 2*n* alleles are randomly permuted the number of times indicated in the PERMS= 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.

Example:

ComputingLinkage Disequilibrium Measures for SNP Data

The data set contains 44 individuals' genotypes at five SNPs



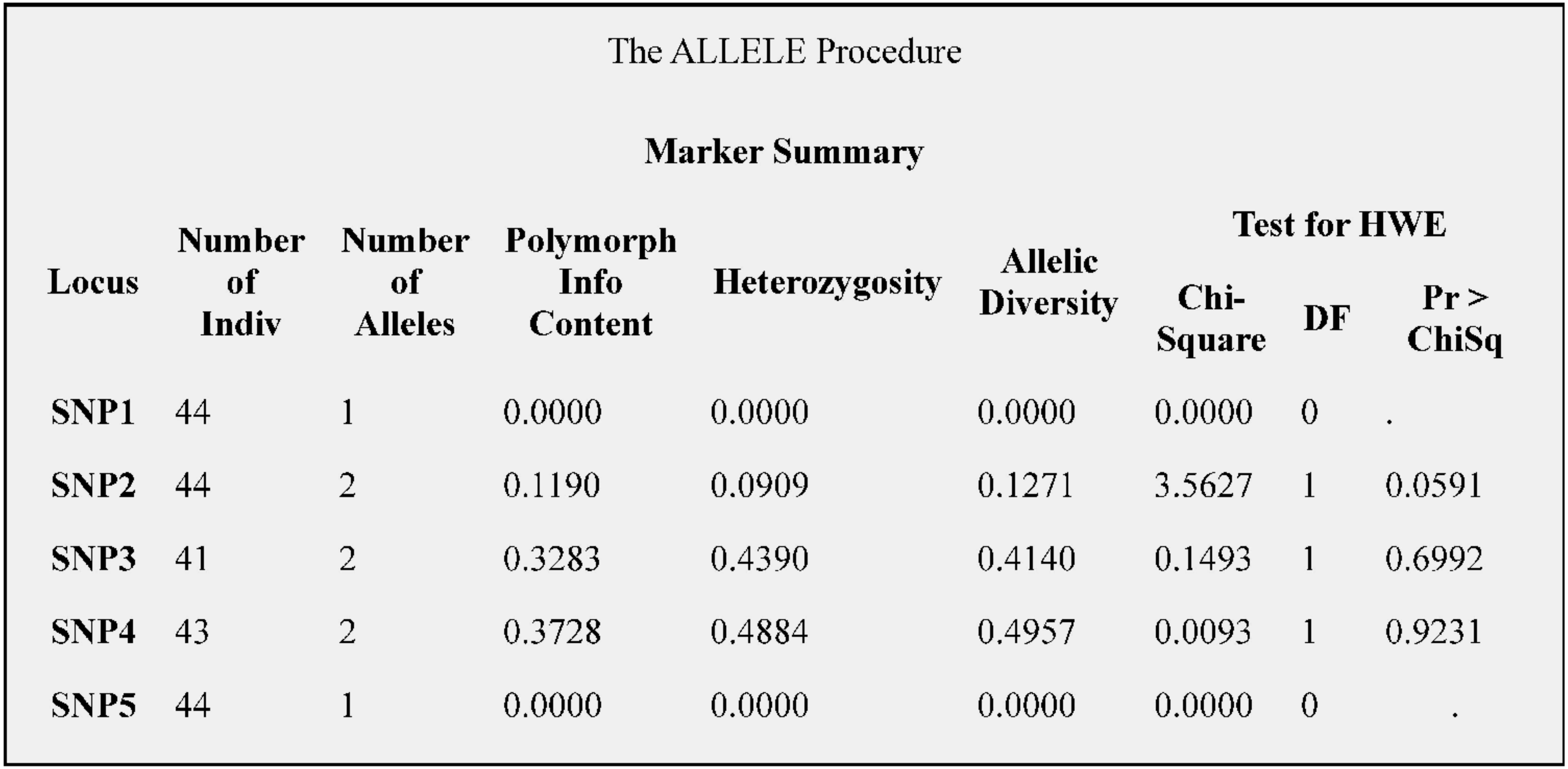
```
data snps;
     input s1-s10;
     datalines;
   2 2 2 1 2 1 1 1 2 2
   2 2 2 2 2 1 1 1 2 2
   2 2 2 2 2 1 2 1 2 2
   2 2 2 2 . . 1 1 2 2
   2 2 2 2 1 2 1 2 2 2
   2 2 2 2 . . 2 1 2 2
   2 2 2 2 2 1 2 1 2 2
   2 2 2 2 . . 2 1 2 2
   2 2 2 2 1 1 1 1 2 2
   2 2 1 1 2 2 2 1 2 2
   2 2 2 1 2 2 2 1 2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

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.

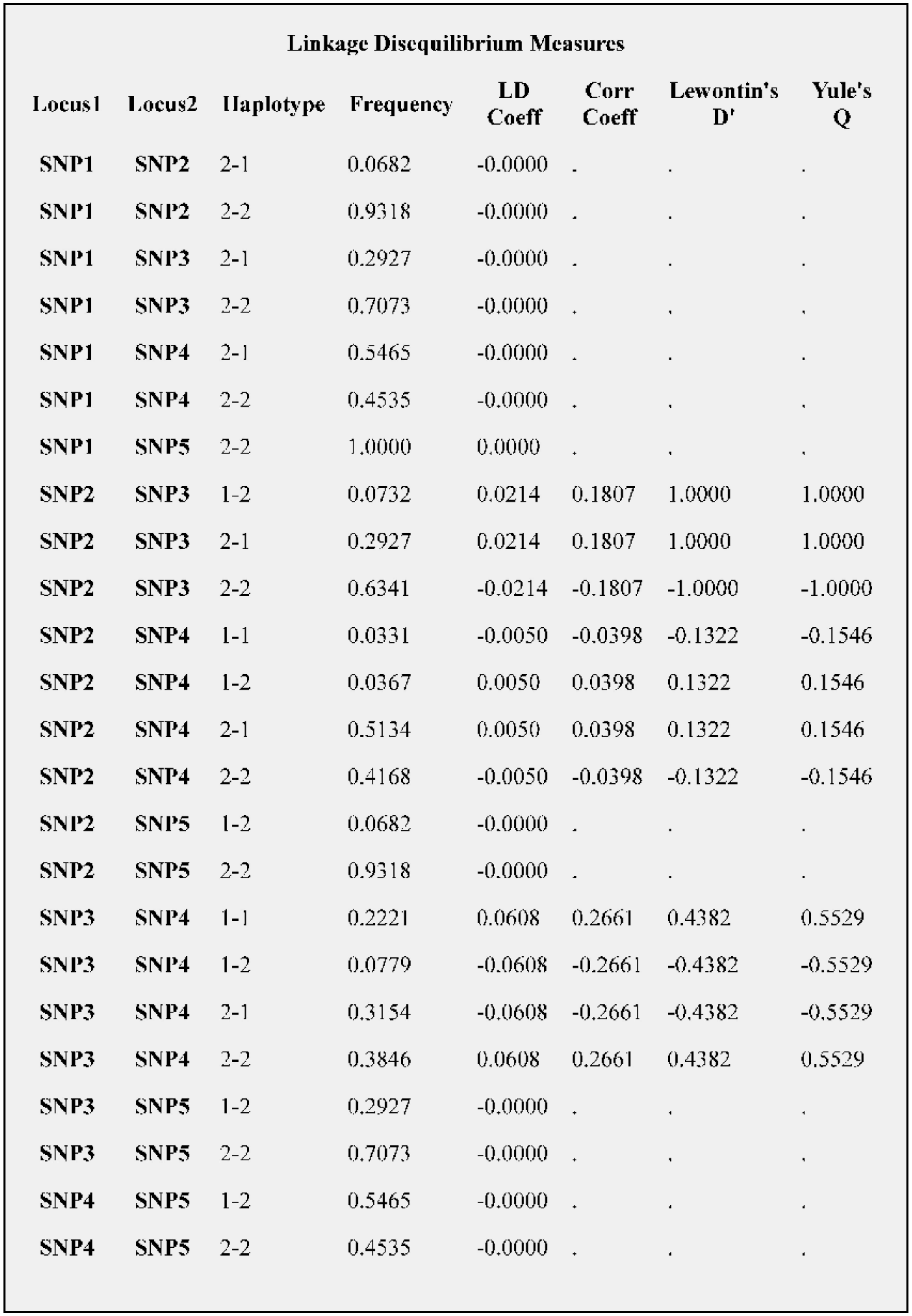


The ALLELE Procedure

Marker Summary

Locus	Number of Indiv	Number of Alleles	Polymorph Info Content	Heterozygosity	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	.

$$\mathrm{PIC} = 1 - \sum_{u=1}^{k} \tilde{p}_u^2 - \sum_{u=1}^{k-1} \sum_{v=u+1}^{k} 2\tilde{p}_u^2\tilde{p}_v^2 \qquad \mathrm{Het} = 1 - \sum_{u=1}^{k} \tilde{P}_{uu} \qquad \mathrm{Div} = 1 - \sum_{u=1}^{k} \tilde{p}_u^2$$



Linkage Disequilibrium Measures

Locus1	Locus2	Haplotype	Frequency	LD Coeff	Corr Coeff	Lewontin's D'	Yule's Q
SNP1	**SNP2**	2-1	0.0682	-0.0000	.	.	.
SNP1	**SNP2**	2-2	0.9318	-0.0000	.	.	.
SNP1	**SNP3**	2-1	0.2927	-0.0000	.	.	.
SNP1	**SNP3**	2-2	0.7073	-0.0000	.	.	.
SNP1	**SNP4**	2-1	0.5465	-0.0000	.	.	.
SNP1	**SNP4**	2-2	0.4535	-0.0000	.	.	.
SNP1	**SNP5**	2-2	1.0000	0.0000	.	.	.
SNP2	**SNP3**	1-2	0.0732	0.0214	0.1807	1.0000	1.0000
SNP2	**SNP3**	2-1	0.2927	0.0214	0.1807	1.0000	1.0000
SNP2	**SNP3**	2-2	0.6341	-0.0214	-0.1807	-1.0000	-1.0000
SNP2	**SNP4**	1-1	0.0331	-0.0050	-0.0398	-0.1322	-0.1546
SNP2	**SNP4**	1-2	0.0367	0.0050	0.0398	0.1322	0.1546
SNP2	**SNP4**	2-1	0.5134	0.0050	0.0398	0.1322	0.1546
SNP2	**SNP4**	2-2	0.4168	-0.0050	-0.0398	-0.1322	-0.1546
SNP2	**SNP5**	1-2	0.0682	-0.0000	.	.	.
SNP2	**SNP5**	2-2	0.9318	-0.0000	.	.	.
SNP3	**SNP4**	1-1	0.2221	0.0608	0.2661	0.4382	0.5529
SNP3	**SNP4**	1-2	0.0779	-0.0608	-0.2661	-0.4382	-0.5529
SNP3	**SNP4**	2-1	0.3154	-0.0608	-0.2661	-0.4382	-0.5529
SNP3	**SNP4**	2-2	0.3846	0.0608	0.2661	0.4382	0.5529
SNP3	**SNP5**	1-2	0.2927	-0.0000	.	.	.
SNP3	**SNP5**	2-2	0.7073	-0.0000	.	.	.
SNP4	**SNP5**	1-2	0.5465	-0.0000	.	.	.
SNP4	**SNP5**	2-2	0.4535	-0.0000	.	.	.

Linkage Disequilibrium 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 , Lewontin's *D*', the proportional difference *d*, and Yule's *Q*.

$$
\begin{aligned}
r &= \frac{D}{(p_1 p_2 q_1 q_2)^{1/2}} \\
\delta &= \frac{D}{q_1 p_{22}} \\
D' &= \frac{D}{D_{\max}}, D_{\max} = \begin{cases} \min(p_1 q_2, q_1 p_2), & D > 0 \\ \min(p_1 q_1, q_2 p_2), & D < 0 \end{cases} \\
d &= \frac{D}{q_1 q_2} \\
Q &= \frac{D}{p_{11} p_{22} + p_{12} p_{21}}
\end{aligned}
$$

HTSNP procedure

Single nucleotide polymorphisms (SNP), roughly one SNP per 1kb in the human genome, accounts for about 90% of human DNA polymorphism

SNPs over large genomic regions suggest the presence of discrete blocks with limited haplotype diversity punctuated by recombination hot spots

Within each block, because of high LD, some allele(s) may always be coexistent with a particular allele at another locus such that

(1) little haplotype diversity exists in the block, and

(2) not all SNPs will be essential in characterizing the haplotype structure in the block

The most common haplotypes could usually be captured by a small subset of SNPs, termed haplotype tag SNPs (htSNPs) by Johnson et al. (2001).

Selection of such a SNP subset that distinguishes all haplotypes is known as the minimum test set problem.

The search space of choosing k SNPs out of m is $\binom{m}{k} = \frac{m!}{k!(m-k)!}$, for which enumerating all possible k-SNP combinations becomes impractical even for moderate numbers of *m* and *k*. HTSNP procedure implements some heuristic algorithms for fast identification of an optimal subset of SNPs without mining through all possible combinations.

Methods of finding Haplotype Tag SNPs

Incremental Search starts with finding a first marker of maximum locus richness and goes through the remaining markers to find each time that, which maximizes PDE.
Decremental Search operates in an opposite manner.

Iterative Maximization Search (Gouesnard et al. 2001) is a fast algorithm for choosing an optimal *k*-subset from m accessions. It starts from a random selection of *k* markers for which all the core collections of size *k*-1 are tested. The subset with the highest PDE is retained. Among the other *m-k* markers, one that brings the greatest increase in the goodness criterion is selected and a new k-locus set is obtained. Exclusion and inclusion of one marker in the new k-locus set is repeated until convergence. Each iteration needs to evaluate the PDE k times for k-1 markers and m-k times for k markers.

Simulated Annealing Search (Kirkpatrick, Gelatt, and Vecchi 1983) has been adopted in many combinatorial optimization problems. /global optimization problem of applied mathematics, namely locating a good approximation to the global minimum of a given function in a large search space. It is often used when the search space is discrete (e.g., all tours that visit a given set of cities). For certain problems, simulated annealing may be more effective than exhaustive enumeration — provided that the goal is merely to find an acceptably good solution in a fixed amount of time, rather than the best possible solution./ Starting from a selection of *k* markers, one marker is randomly swapped with another from the unselected markers. The change of haplotype goodness is evaluated using an energy function for the marker exchange. Acceptance of the exchange is judged with the Metropolis criterion (Metropolis et al. 1953) using the change *Δ* of energy function and the annealing temperature *T*:

$$\Pr\{\text{new point is accepted}\} = \begin{cases} 1, & \Delta \leq 0 \\ \exp(-\Delta/T), & \Delta > 0 \end{cases}$$

Exhaustive Search An exhaustive search of k markers from m involves traversal of all possible selections once and only once.

Calculations

For *n* haplotypes diversity D_H records the weighted differences of all n^2 pairwise comparisons of two haplotypes (Clayton 2002).

$$D_H = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j (h_i - h_j)$$

Here h_i and h_j denote the *i*-th and the *j*-th haplotype, and w_i and w_j are the corresponding weights. The difference of two *m*-locus haplotypes, h_i and h_j, is computed as the total allele differences at the *m* loci:

$$h_i - h_j = \sum_{k=1}^{m} (h_{ik} - h_{jk})$$

where h_{ik} is the allele of the *i*-th haplotype observed at the *k*-th locus and

$$h_{ik} - h_{jk} = \begin{cases} 0 & h_{ik} = h_{jk} \\ 1 & h_{ik} \neq h_{jk} \end{cases}$$

If only distinct haplotypes are recorded with their corresponding frequencies using the FREQ statement, then F_i, the weighted frequency of haplotype h_i, can be calculated as

$$F_i = \frac{w_i f_i}{\sum_j w_j f_j}$$

where w_j and f_j are the weight and frequency. The estimate of haplotype diversity D_H is then proportional to the average of the gene diversity D_k at all *m* loci, computed as

$$D = \frac{\sum_{k=1}^{m} D_k}{m}$$

where $D_k = 1 - \sum_{u=1}^{l_k} p_{ku}^2$ and l_k is the number of alleles at the *k*-th locus. The weighted allele frequency p_{ku} of the *u*-th allele at the *k*-th locus is recorded as

$$p_{ku} = \sum_{h_{ik}=u} F_i$$

The diversity computed in this way measures the probability that two haplotypes sampled from the population differ at any locus.

• *Goodness of a Core Set Selection*

For a core set of *k* SNPs, the *n* observed haplotypes can be classified into *G* groups of *k*-locus haplotypes. The haplotype residual diversity, R_D, is defined as the sum of the within-group diversities for the *G* groups: ------------->

$$R_D = \sum_{g=1}^{G} F_g^2 \Big[\sum_{i \in G_g} \sum_{j \in G_g} F_{ig} F_{jg} (h_i - h_j) \Big]$$

where F_g is the weighted frequency of the *g*-th group calculated as ------------->

$$F_g = \textstyle\sum_{i \in G_g} F_i$$

and F_{ig} is the within-group weighted frequency for the *i*-th haplotype in the *g*-th group with $F_{ig} = F_i/F_g$.

Similarly, R_D can be calculated as the total within-group gene diversity: ------------->

$$R_D = \sum_{g=1}^{G} F_g^2 D_g$$

and ------------->

$$D_g = \frac{\sum_{k=1}^{m} D_{kg}}{m}$$

where ------------->

$$D_{kg} = 1 - \textstyle\sum_u p_{kug}^2$$

is calculated using the within-group allele frequencies $p_{k\,u\,g}$.

The proportion of diversity explained (PDE) by a SNP set selection is used to evaluate the goodness of that selection. *PDE* is calculated as

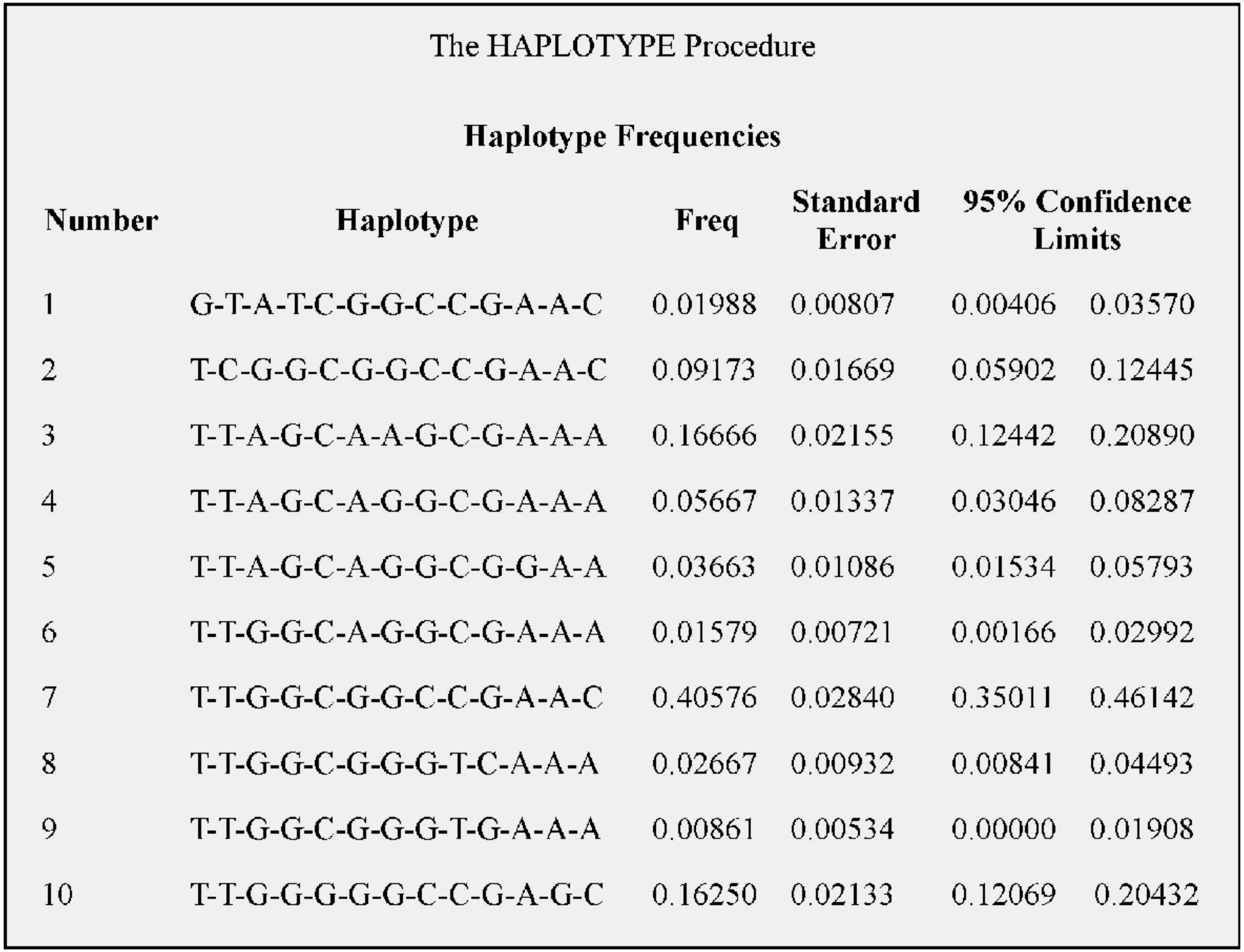
------------->

$$PDE = 1-[(R_D)/D]$$

The selected search algorithm finds the optimal subset that maximizes *PDE*.

Using the HAPLOTYPE and HTSNP Procedures Together

Before using PROC HTSNP, you may need to first run PROC HAPLOTYPE if you have data with unknown phase in order to estimate the haplotype frequencies



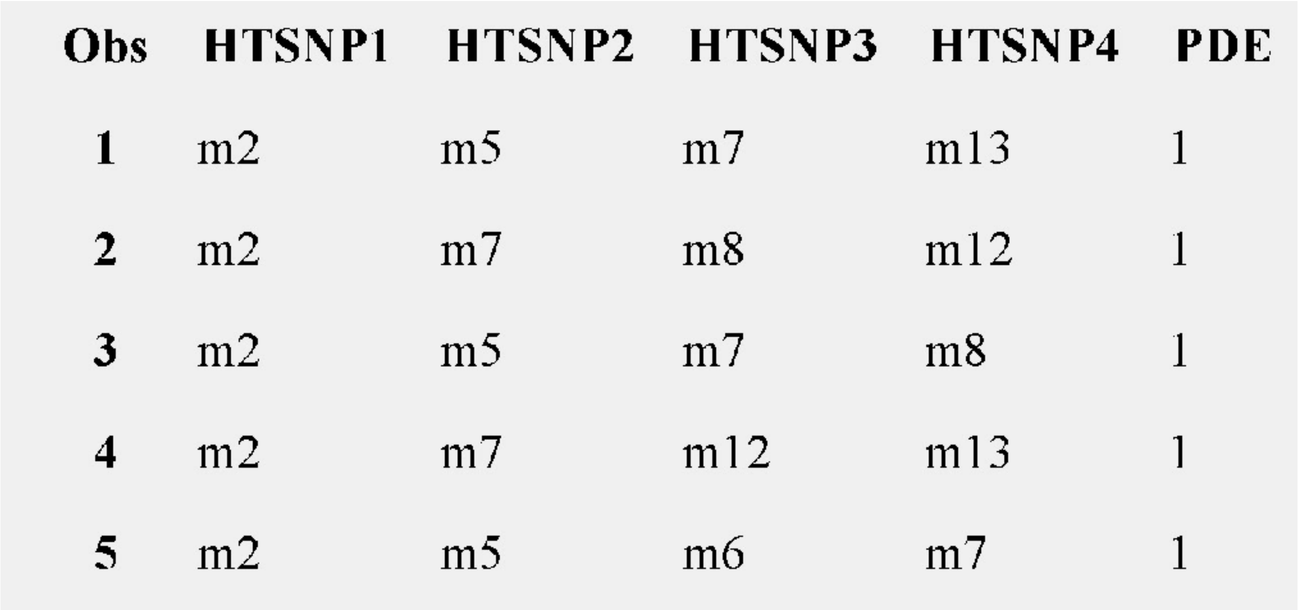
The HAPLOTYPE Procedure

Haplotype Frequencies

Number	Haplotype	Freq	Standard Error	95% Confidence Limits	
1	G-T-A-T-C-G-G-C-C-G-A-A-C	0.01988	0.00807	0.00406	0.03570
2	T-C-G-G-C-G-G-C-C-G-A-A-C	0.09173	0.01669	0.05902	0.12445
3	T-T-A-G-C-A-A-G-C-G-A-A-A	0.16666	0.02155	0.12442	0.20890
4	T-T-A-G-C-A-G-G-C-G-A-A-A	0.05667	0.01337	0.03046	0.08287
5	T-T-A-G-C-A-G-G-C-G-G-A-A	0.03663	0.01086	0.01534	0.05793
6	T-T-G-G-C-A-G-G-C-G-A-A-A	0.01579	0.00721	0.00166	0.02992
7	T-T-G-G-C-G-G-C-C-G-A-A-C	0.40576	0.02840	0.35011	0.46142
8	T-T-G-G-C-G-G-G-T-C-A-A-A	0.02667	0.00932	0.00841	0.04493
9	T-T-G-G-C-G-G-G-T-G-A-A-A	0.00861	0.00534	0.00000	0.01908
10	T-T-G-G-G-G-G-C-C-G-A-G-C	0.16250	0.02133	0.12069	0.20432

```
proc htsnp data=hapfreq
     Size=4               /* Number of tag SNPs */
     method=sa            /* Simulated Annealing Method */
     Best=5               /* Five best SNPs tag sets */
     cutoff=0.05
     Seed=123             /* Initializing the random selection of first the set */
     outstat=out;
     var m1-m13;          /* SNPs in haplotypes */
     freq freq;           /* Weights of different haplotypes */
run;
```

In this example, the Simulated Annealing Search method is specified for finding the best sets of size four. The output data set OUT that is created by PROC HTSNP is then printed out to show the best five sets of SNPs that were selected.



Obs	HTSNP1	HTSNP2	HTSNP3	HTSNP4	PDE
1	m2	m5	m7	m13	1
2	m2	m7	m8	m12	1
3	m2	m5	m7	m8	1
4	m2	m7	m12	m13	1
5	m2	m5	m6	m7	1

HAPLOTYPE procedure

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.

PROC HAPLOTYPE performs a likelihood ratio test to test the hypothesis of no LD between marker loci.

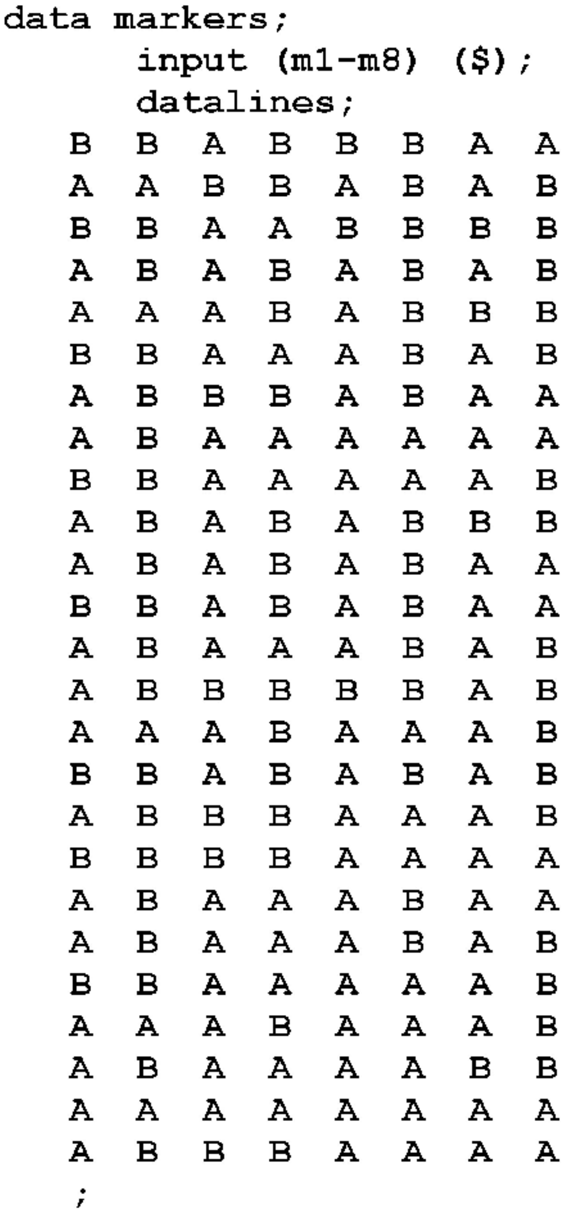
Another application is association testing of disease susceptibility: haplotypes might include two or more causative sites show synergistic interaction.

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.

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

Example



```
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
   ;
```

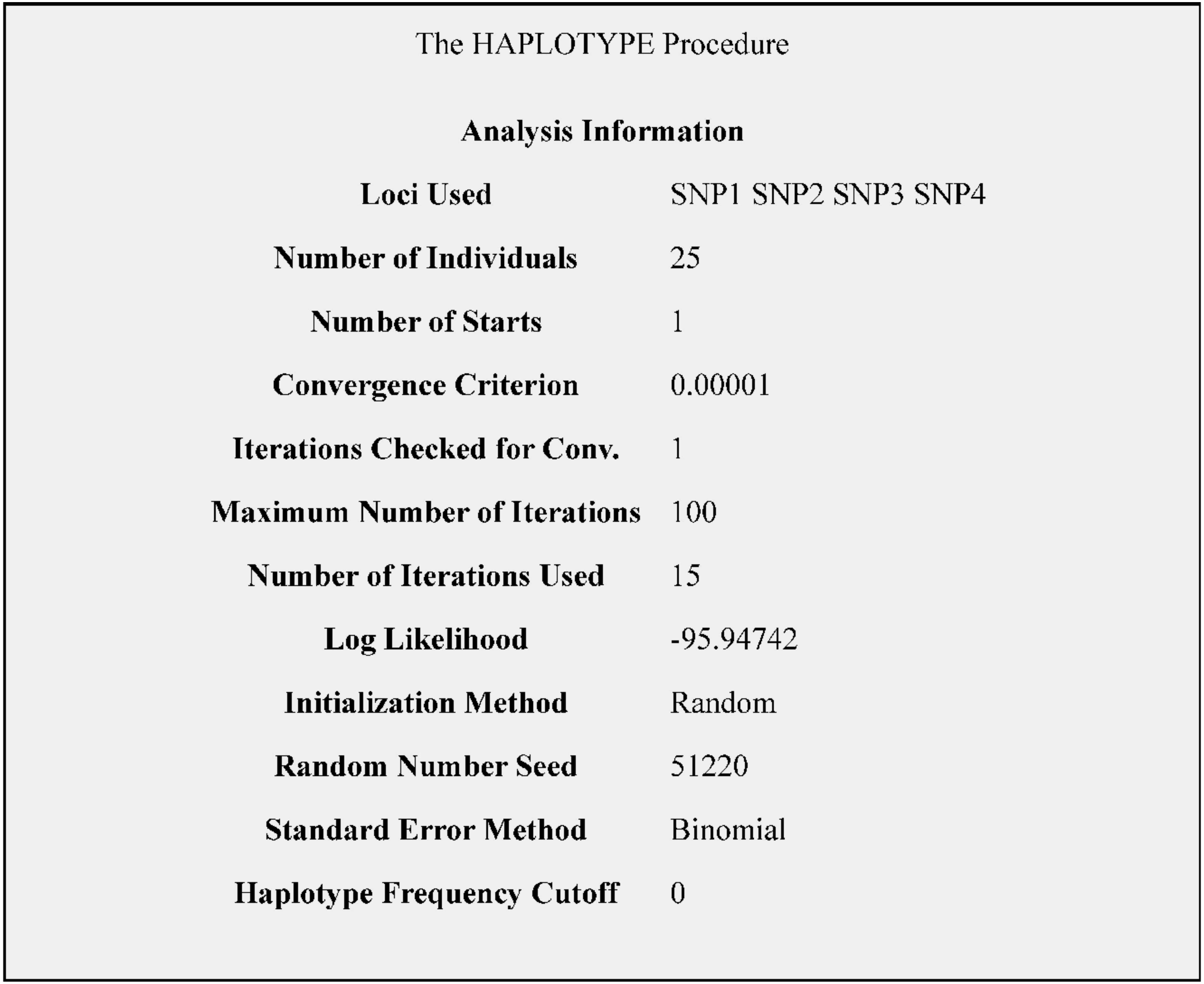
You can now use PROC HAPLOTYPE to infer the possible haplotypes and estimate the four-locus haplotype frequencies in this sample. The following statements will perform these calculations:

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

EM algorithm estimates the haplotype frequencies

Option INIT=RANDOM indicates that initial haplotype frequencies are randomly generated

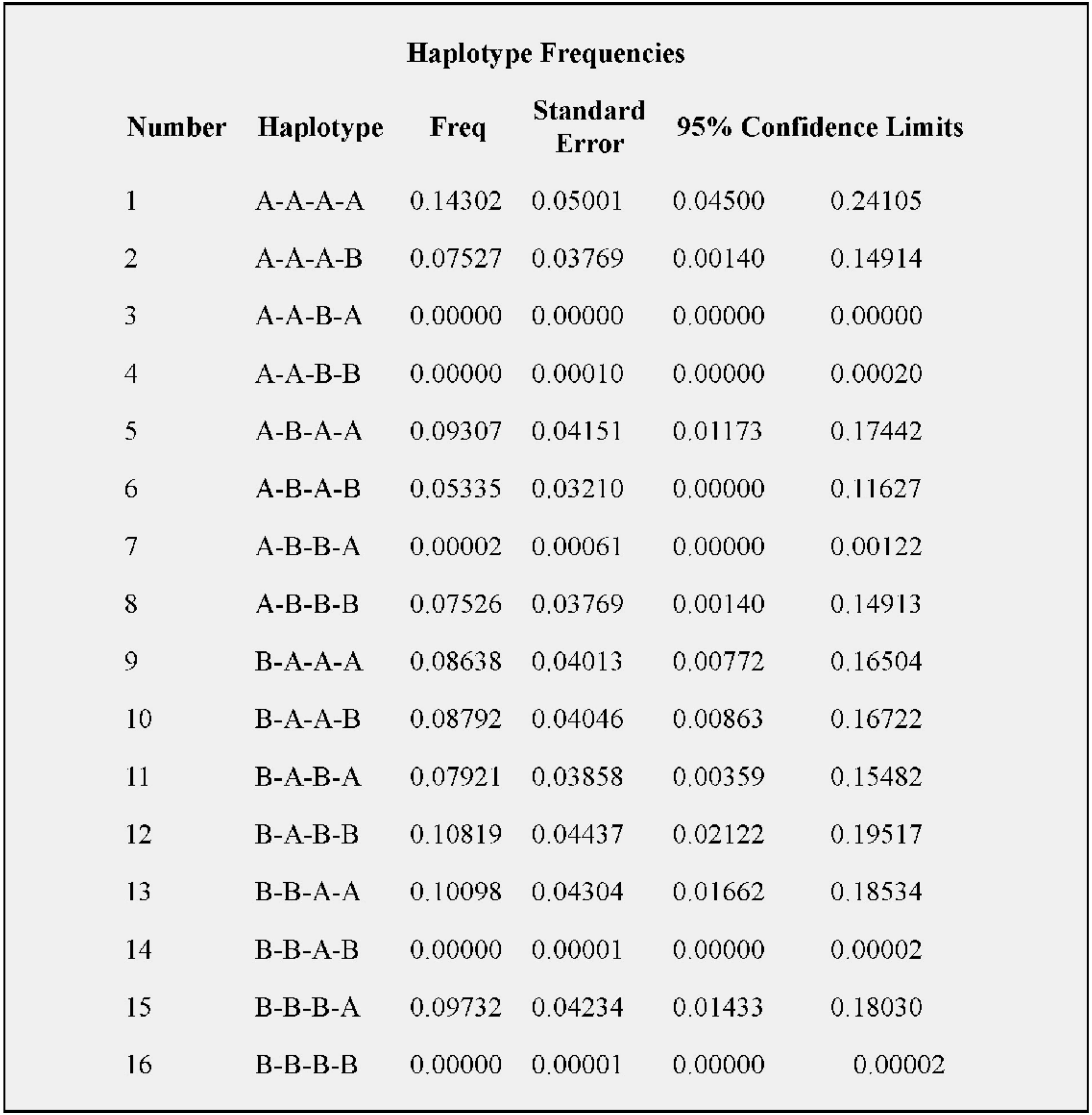
Maximum number of iterations is set to 100.



```
                    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
```

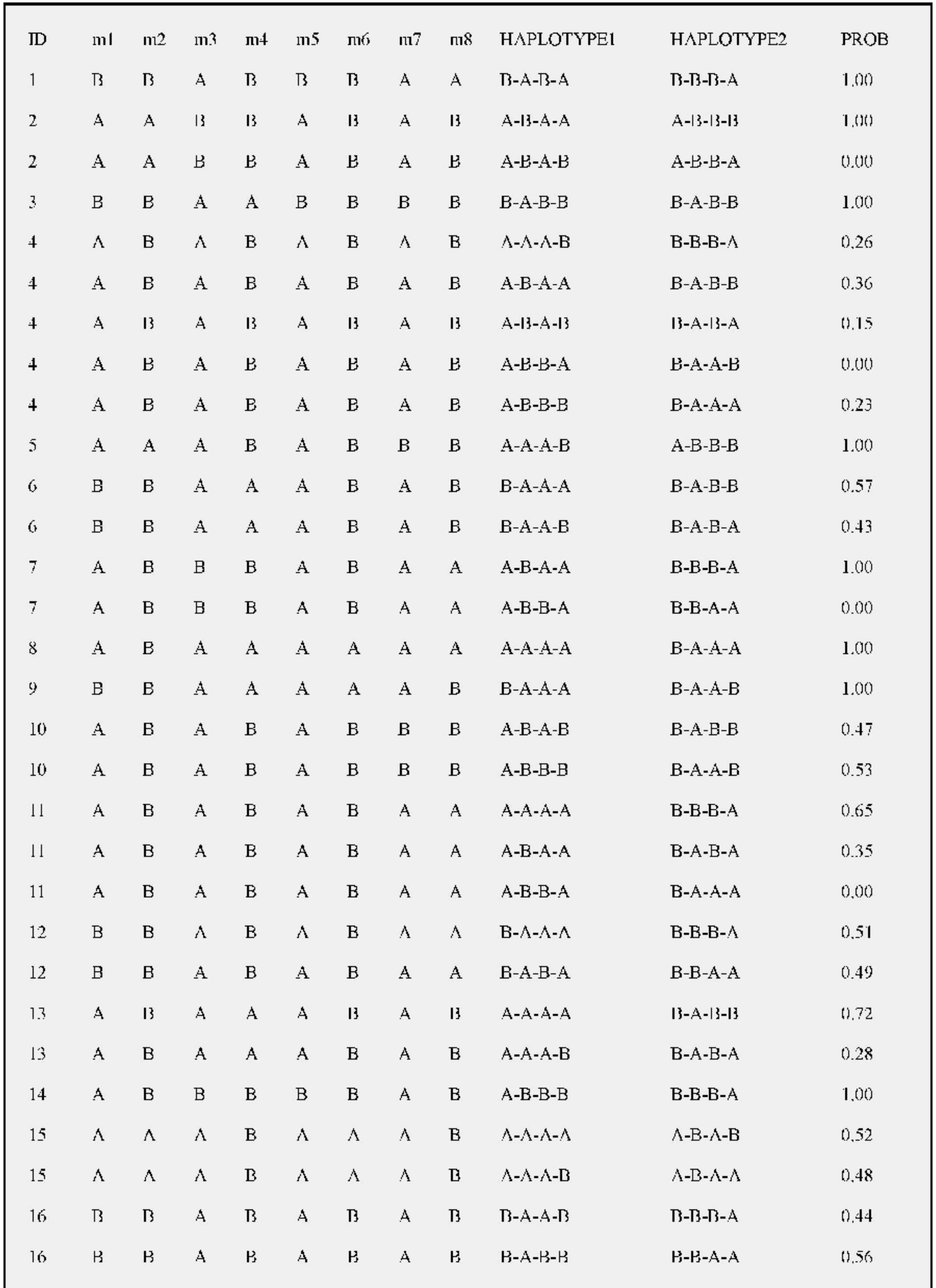


Haplotype Frequencies

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

Output Data Set from the HAPLOTYPE Procedure

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.

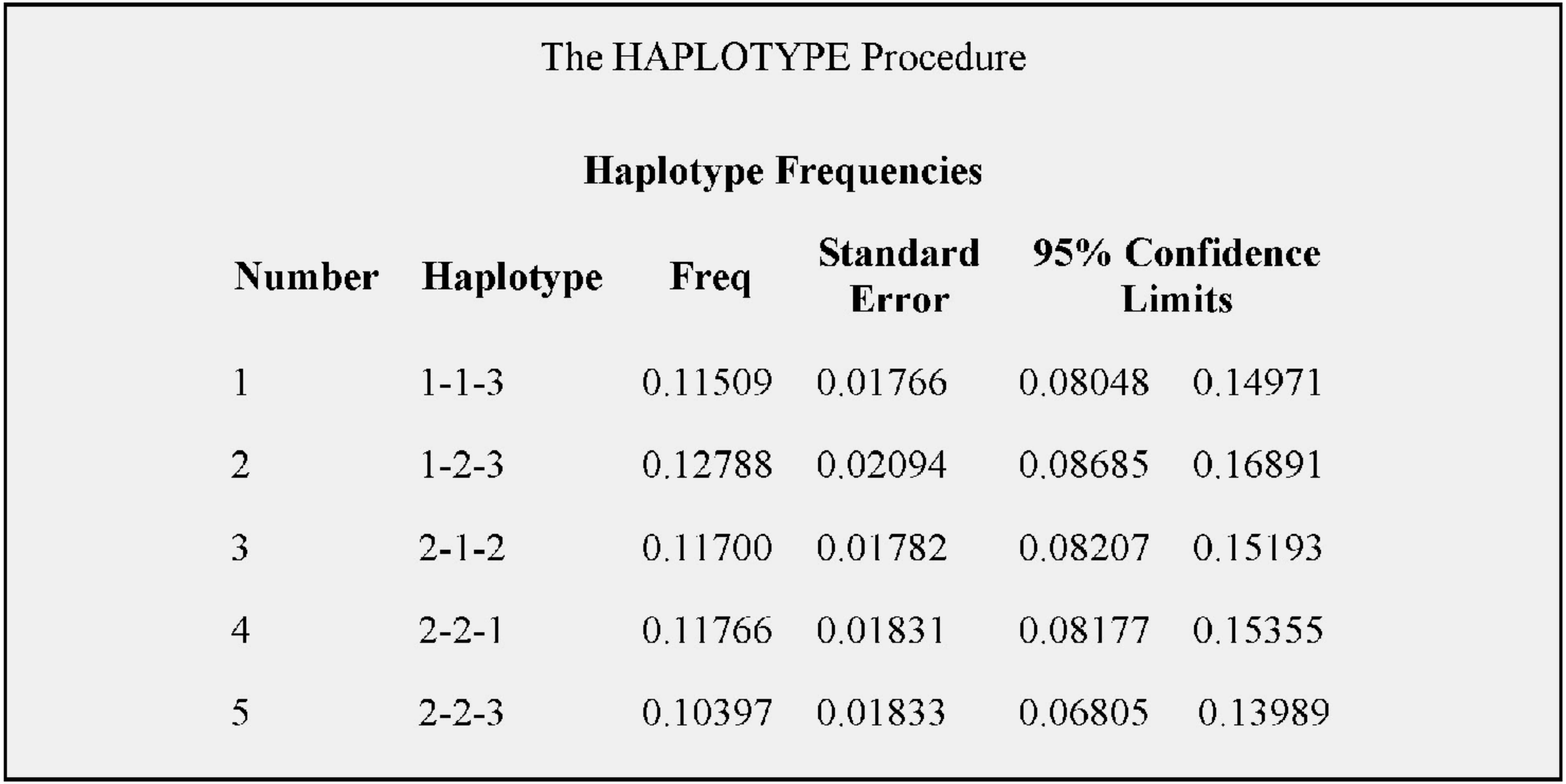


ID	m1	m2	m3	m4	m5	m6	m7	m8	HAPLOTYPE1	HAPLOTYPE2	PROB
1	B	B	A	B	B	B	A	A	B-A-B-A	B-B-B-A	1.00
2	A	A	B	B	A	B	A	B	A-B-A-A	A-B-B-B	1.00
2	A	A	B	B	A	B	A	B	A-B-A-B	A-B-B-A	0.00
3	B	B	A	A	B	B	B	B	B-A-B-B	B-A-B-B	1.00
4	A	B	A	B	A	B	A	B	A-A-A-B	B-B-B-A	0.26
4	A	B	A	B	A	B	A	B	A-B-A-A	B-A-B-B	0.36
4	A	B	A	B	A	B	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

CUTOFF= option affects the "Haplotype Frequencies" table. To view only the haplotypes with an estimated frequency of at least 0.10:

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

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



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

- References

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

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.

Schaid, D.J., Rowland, C.M., Tines, D.E., Jacobson, R.M., and Poland, G.A. (2002), "Score Tests for Association between Traits and Haplotypes when Linkage Phase is Ambiguous," *American Journal of Human Genetics,* 70, 425 -434.

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," *Human Heredity,* 53, 79 -91.

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

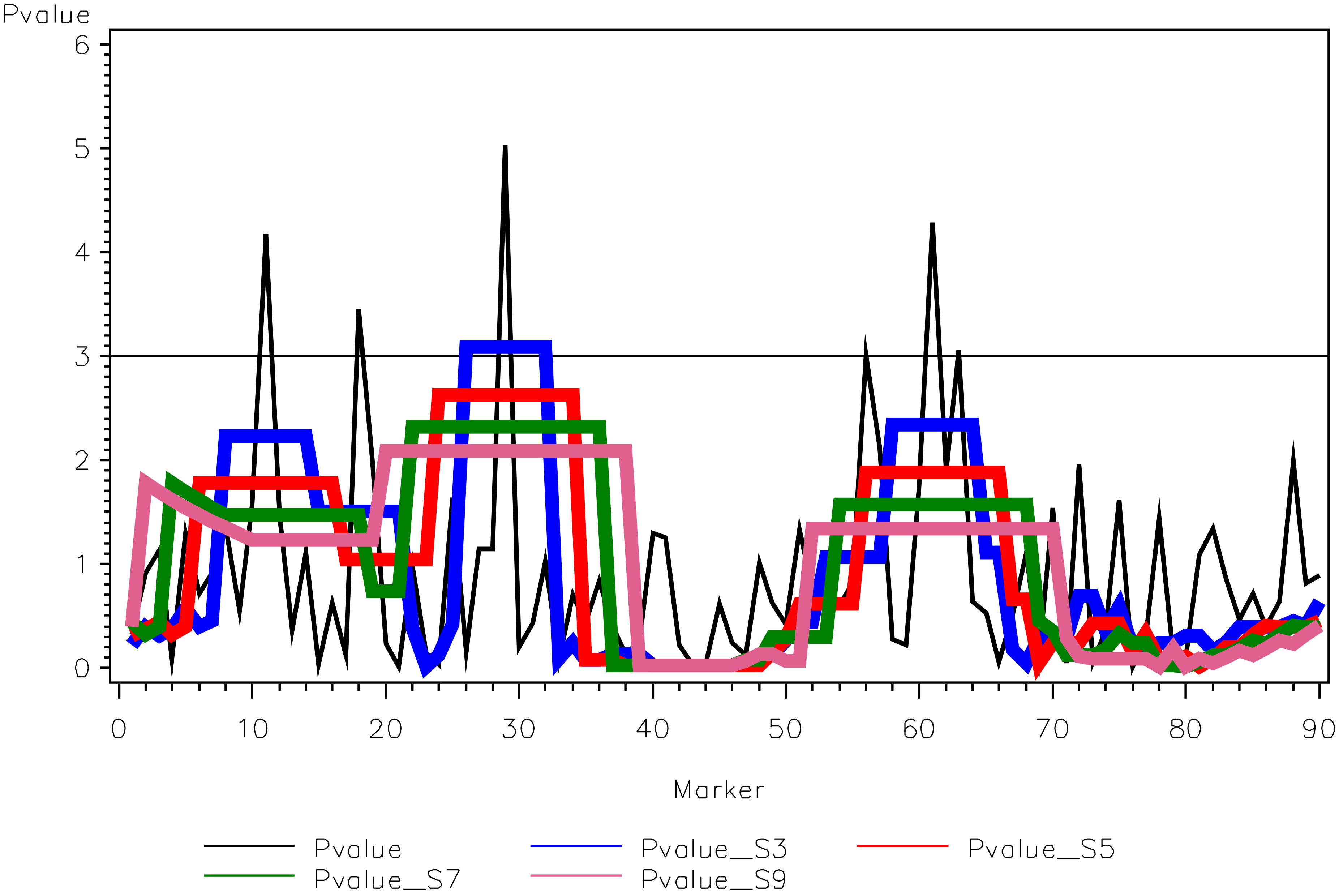
PSMOOTH procedure

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 Simes' method (1986), Fisher's method (1932), and/or the truncated product method (TPM) (2002) 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).

```
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
;
```

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


Pvalue
6
5
4
3
2
1
0
0
10
20
30
40
50
60
70
80
90
Marker
Pvalue
Pvalue__S3
Pvalue__S5
Pvalue__S7
Pvalue__S9

CASECONTROL procedure

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.

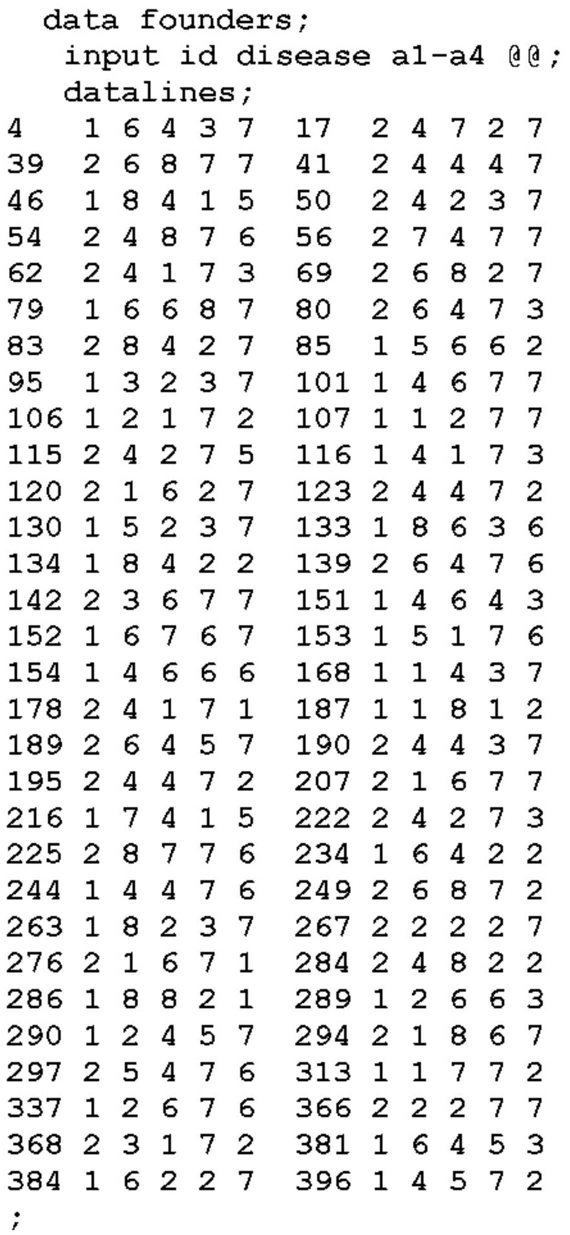
The three chi-square tests available for testing case-control genotypic data are the genotype case-control test, which tests for dominant allele effects on the disease penetrance, and the allele case-control test and linear trend test, which test for additive allele effects on the disease penetrance. 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.

OUTSTAT= Data Set

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

- BY variables, if any
- Locus
- Counts of genotyped individuals for the two values of the TRAIT variable: NumTrait1 and NumTrait2, where 1 and 2 are replaced by the values of the TRAIT variable
- Chi-square statistic for each test performed: ChiSqAllele, ChiSqGenotype, and ChiSqTrend
- Degrees of freedom for each test performed: dfAllele, dfGenotype, and dfTrend
- *p*-value for each test performed: ProbAllele, ProbGenotype, and ProbTrend

Example



```
  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
;
```

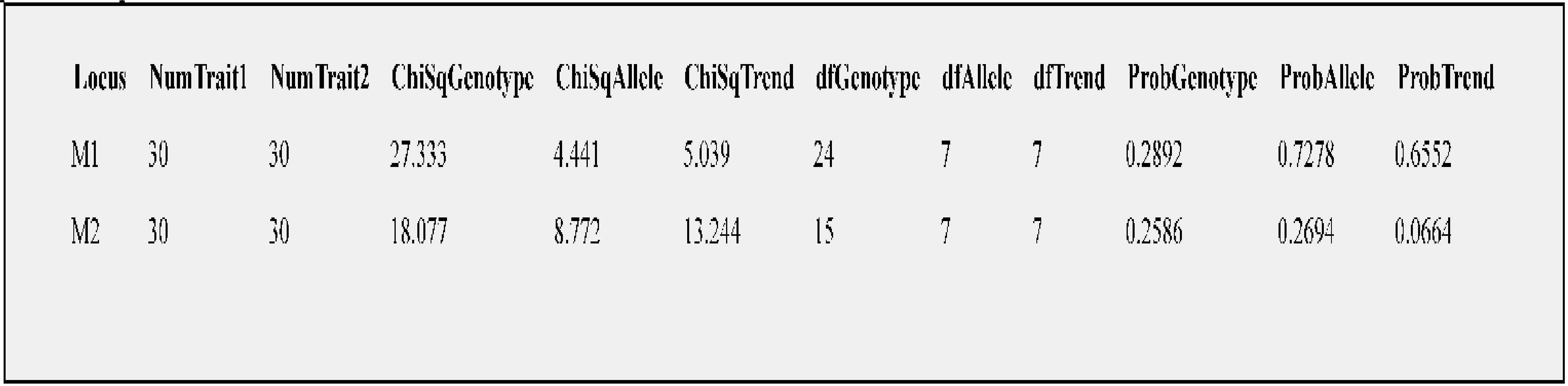
The multiallelic versions of the association tests are performed since each marker has more than two alleles.

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.

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

proc print noobs heading=h; run;
```

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



```
Locus  NumTrait1  NumTrait2  ChiSqGenotype  ChiSqAllele  ChiSqTrend  dfGenotype  dfAllele  dfTrend  ProbGenotype  ProbAllele  ProbTrend

M1     30         30         27.333         4.441        5.039       24          7         7        0.2892        0.7278      0.6552

M2     30         30         18.077         8.772        13.244      15          7         7        0.2586        0.2694      0.0664
```

INBREED procedure

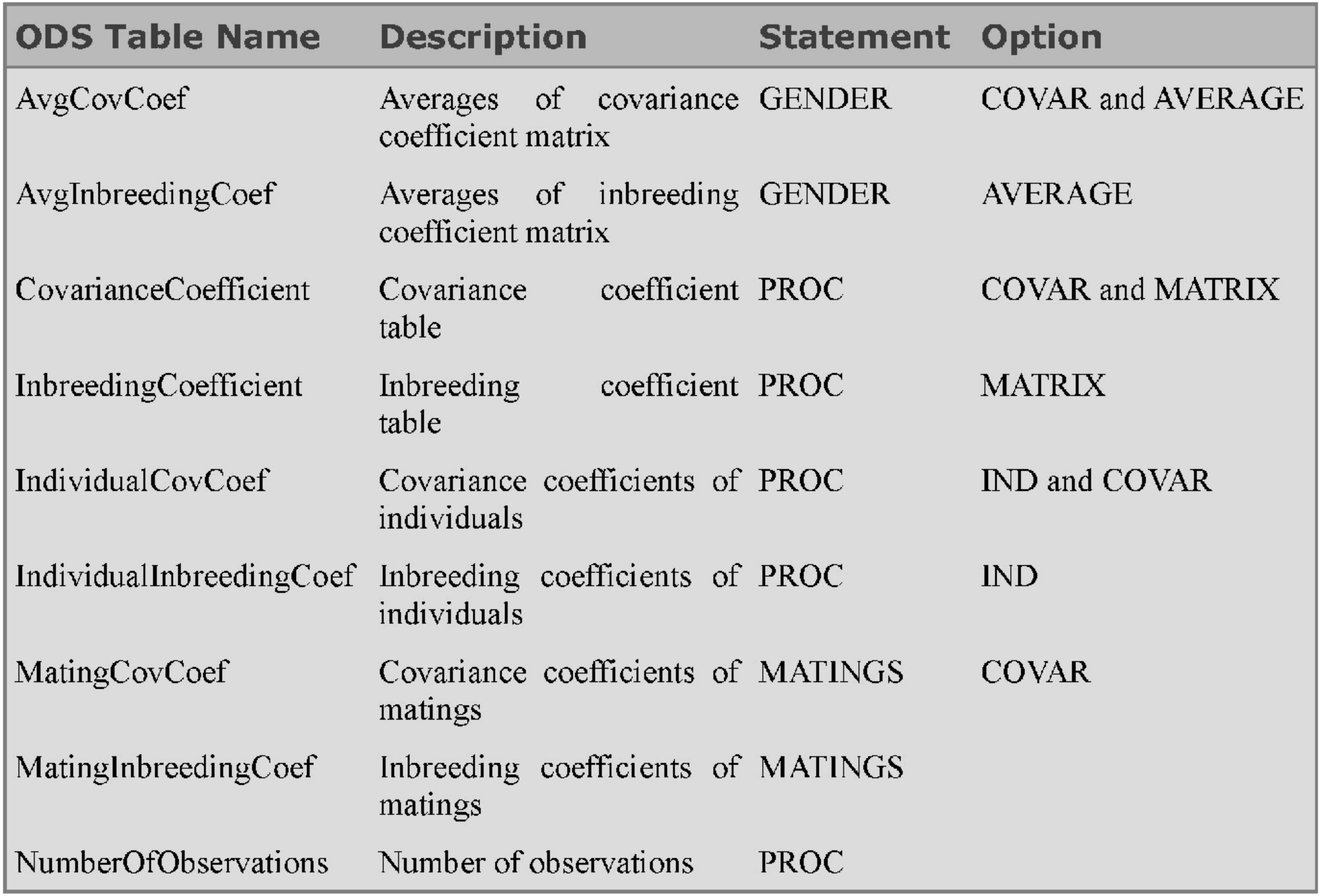
The INBREED procedure calculates the covariance or inbreeding coefficients for a pedigree. PROC INBREED is unique in that it handles very large populations.

The INBREED procedure has two modes of operation. One mode carries out analysis on the assumption that all the individuals belong to the same generation.

The other mode divides the population into nonoverlapping generations and analyzes each generation separately, assuming that the parents of individuals in the current generation are defined in the previous generation.

PROC INBREED also computes averages of the covariance or inbreeding coefficients within sex categories if the sex of individuals is known.

ODS Tables Produced in PROC INBREED



ODS Table Name	Description	Statement	Option
AvgCovCoef	Averages of covariance coefficient matrix	GENDER	COVAR and AVERAGE
AvgInbreedingCoef	Averages of inbreeding coefficient matrix	GENDER	AVERAGE
CovarianceCoefficient	Covariance coefficient table	PROC	COVAR and MATRIX
InbreedingCoefficient	Inbreeding coefficient table	PROC	MATRIX
IndividualCovCoef	Covariance coefficients of individuals	PROC	IND and COVAR
IndividualInbreedingCoef	Inbreeding coefficients of individuals	PROC	IND
MatingCovCoef	Covariance coefficients of matings	MATINGS	COVAR
MatingInbreedingCoef	Inbreeding coefficients of matings	MATINGS	
NumberOfObservations	Number of observations	PROC	

FAMILY procedure

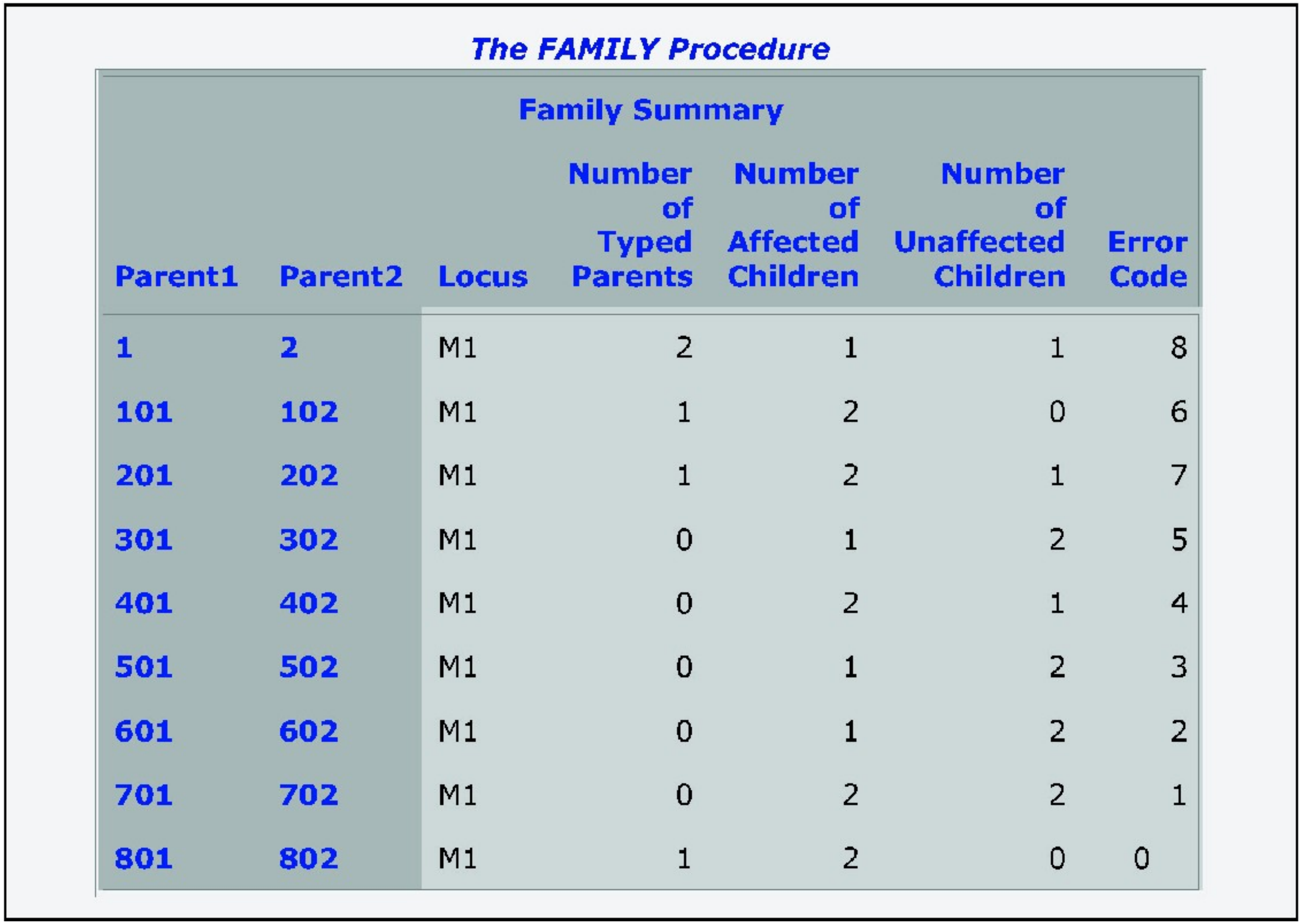
Often provide a more effective way of testing markers for association with disease status than case-control data. Case-control data may uncover significant associations between markers and a disease that could be caused by factors other than linkage, such as population structure.

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).

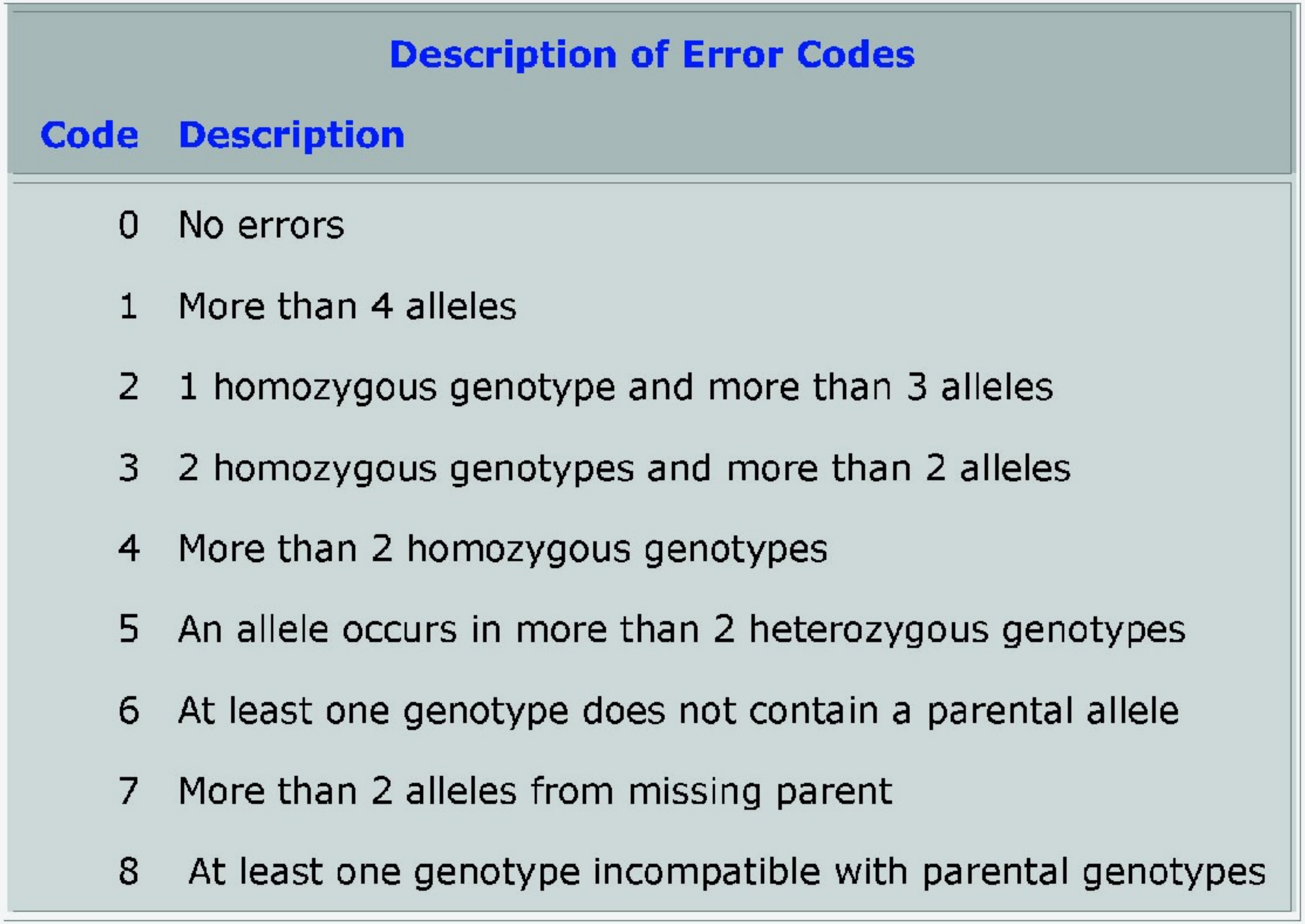
Both types of families can be jointly analyzed using the combined versions of the S-TDT and SDT and the reconstruction-combined TDT (RC-TDT). The RC-TDT can additionally accommodate families with no unaffected children and missing parental genotypes in certain situations.



The FAMILY Procedure

Family Summary

Parent1	Parent2	Locus	Number of Typed Parents	Number of Affected Children	Number of Unaffected Children	Error Code
1	2	M1	2	1	1	8
101	102	M1	1	2	0	6
201	202	M1	1	2	1	7
301	302	M1	0	1	2	5
401	402	M1	0	2	1	4
501	502	M1	0	1	2	3
601	602	M1	0	1	2	2
701	702	M1	0	2	2	1
801	802	M1	1	2	0	0



Description of Error Codes

Code	Description
0	No errors
1	More than 4 alleles
2	1 homozygous genotype and more than 3 alleles
3	2 homozygous genotypes and more than 2 alleles
4	More than 2 homozygous genotypes
5	An allele occurs in more than 2 heterozygous genotypes
6	At least one genotype does not contain a parental allele
7	More than 2 alleles from missing parent
8	At least one genotype incompatible with parental genotypes

TPLOT macro

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

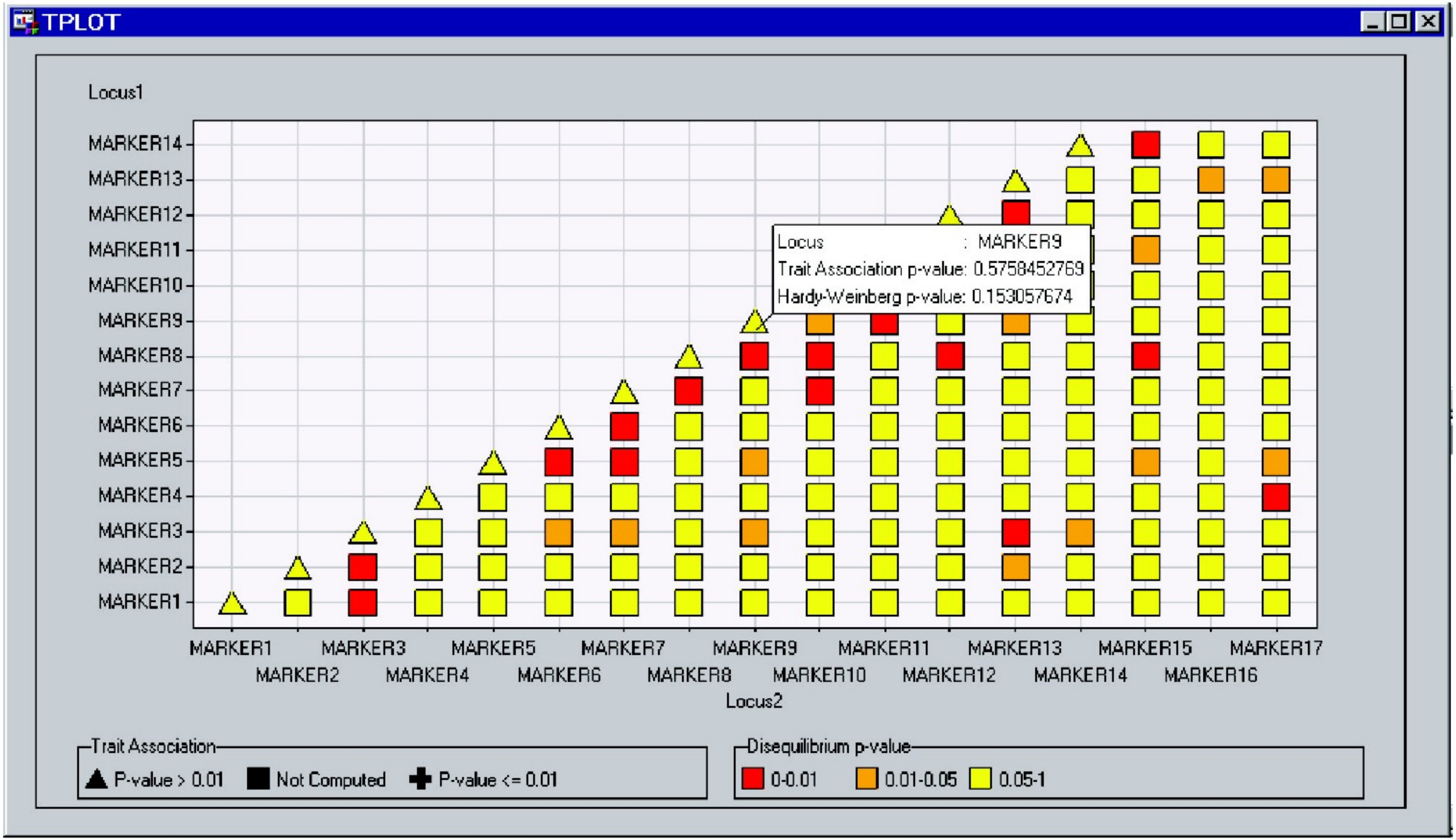
associations between markers and a dichotomous trait (such as disease status). This is a convenient way of combining information contained in output data sets from two separate SAS/Genetics procedures and summarizing it in an easily interpretable plot.

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.

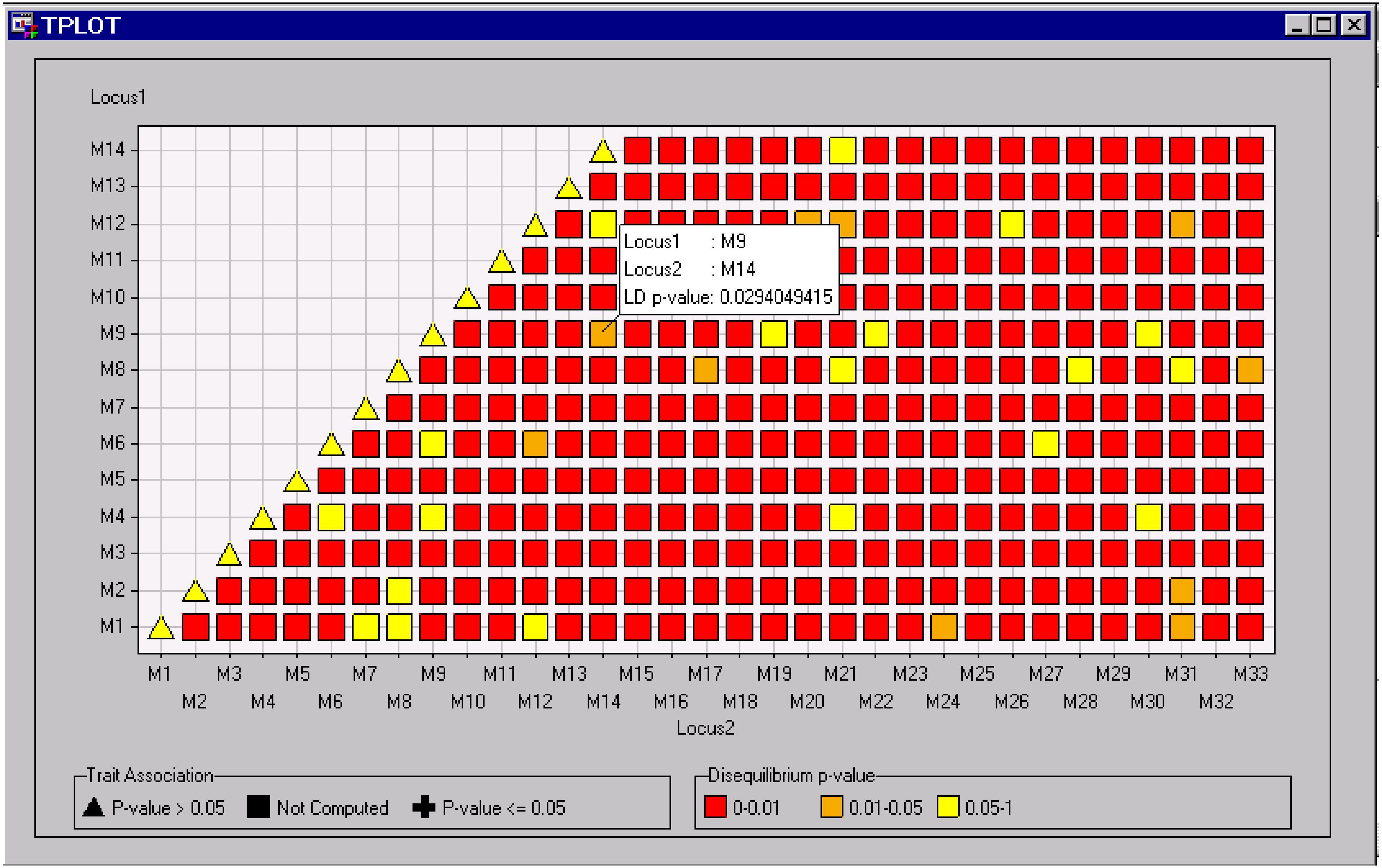
The %TPLOT macro is a part of the SAS Autocall library, and is automatically available for use in your SAS program provided that the SAS system option MAUTOSOURCE is in effect.

Colors and shapes of the data points are used to symbolize *p*-value ranges.
The button in the toolbar enables the *p*-values to be displayed.

Results Window for TPLOT Macro



Locus : MARKER9
Trait Association p-value: 0.5758452769
Hardy-Weinberg p-value: 0.153057674



```
Locus1      : M9
Locus2      : M14
LD p-value: 0.0294049415
```

Aitäh kuulamast !