Association Studies for Non-Family Data

What is allelic association?
“The excessive co-occurrence of certain combinations of alleles in the same gamete because of tight linkage, or for other reasons, is known as allelic association.” 1

Consider two loci A and B, with alleles A₁, A₂, ..., Aₘ and B₁, B₂, ..., Bₙ occurring at frequencies p₁, p₂, ..., pₘ and q₁, q₂, ..., qₙ in the population. We can consider an individual’s two haplotypes with respect to these loci, one of maternal origin and one of paternal origin.

How many possible haplotypes are there?

mn

The haplotypes can be denoted A₁B₁, A₁B₂, ..., AₘBₙ, with frequencies h₁₁, h₁₂, ..., hₘₙ.

If the occurrence of allele Aᵢ and the occurrence of allele Bⱼ in a haplotype are independent events, then, by definition hᵢⱼ = pᵢqⱼ.

“independent alleles” hᵢⱼ = pᵢqⱼ.
“positively associated alleles” hᵢⱼ > pᵢqⱼ
“negatively associated alleles” hᵢⱼ < pᵢqⱼ

Recall our definition of the linkage disequilibrium coefficient from handout 3:

Dᵢⱼ = hᵢⱼ − pᵢqⱼ.

1 Pak Sham, Statistics in Human Genetics, Arnold 1998, p. 145. Much of the material in this handout is taken from this text.
Exercise: Consider two biallelic loci A and B. There are four possible haplotypes: A_1B_1, A_1B_2, A_2B_1, A_2B_2. Suppose that the frequencies of these four haplotypes in a large population are 0.4, 0.1, 0.2, and 0.3, respectively. Are there any allelic associations between these loci? What are they?

Allele frequencies:

A_1: 0.4 + 0.1 = 0.5
A_2: 0.2 + 0.3 = 0.5
B_1: 0.4 + 0.2 = 0.6
B_2: 0.1 + 0.3 = 0.4

If there were no allelic associations, we should have:
P(A_1B_1) = (0.5)(0.6) = 0.3 < 0.4 – positive association
P(A_1B_2) = (0.5)(0.4) = 0.2 > 0.1 – negative association
P(A_2B_1) = (0.5)(0.6) = 0.3 > 0.2 – negative association
P(A_2B_2) = (0.5)(0.4) = 0.2 < 0.3 – positive association
How are allelic associations generated?

- random genetic drift
- founder effects
- mutation
- selection
- population stratification

Genetic drift: In a finite population, the gene pool of one generation can be regarded as a random sample of the gene pool of the previous generation. As such, allele and haplotypes frequencies are subject to sampling variation – random chance. The smaller the population is, the larger the effects of genetic drift are.

Mutation: If a new mutation appears in a population, alleles at loci linked with the mutant allele will maintain an association for many generations. The association lasts longer when linkage is greater (that is, the recombination fraction is much smaller than \( \frac{1}{2} \) – very close to 0).

Founder effects: Applies to a population that has grown rapidly from a small group of ancestors. For example, the 5,000,000 Finns mostly descended from about 1000 people who lived about 2000 years ago. Such a population is prone to allelic disequilibrium due to random genetic drift.

Selection: When an individual’s genotype influences his/her reproductive fitness. For example, if two alleles interact to decrease reproductive fitness, the alleles will tend to be negatively associated.
Stratification: Some populations consist of two or more subgroups that, for cultural or other reasons, have evolved more or less separately. Two loci that are in linkage equilibrium for each subpopulation may be in linkage disequilibrium for the larger population. The following example is a population with three subpopulations. Consider two biallelic loci, the first with alleles A and a and the second with alleles B and b.

<table>
<thead>
<tr>
<th>N</th>
<th>A allele frequency</th>
<th>B allele frequency</th>
<th>AB haplotype frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>0.3</td>
<td>0.5</td>
<td>0.15</td>
</tr>
<tr>
<td>2000</td>
<td>0.2</td>
<td>0.4</td>
<td>0.08</td>
</tr>
<tr>
<td>10000</td>
<td>0.05</td>
<td>0.1</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Do any of three subpopulations show allelic association?

No.  
0.15=0.3x0.5  
0.08=0.2x0.4  
0.005=0.05x0.1

Does the larger population show allelic association?

A allele frequency: \[\frac{1000(0.3)+2000(0.5)+10000(0.1)}{13000}=0.0923\]  
B allele frequency: \[\frac{1000(0.5)+2000(0.4)+10000(0.1)}{13000}=0.1770\]

The equilibrium frequency of AB is (0.0923)(0.1770)=0.0163  
AB haplotype frequency: \[\frac{1000(0.15)+2000(0.08)+10000(0.005)}{13000}=0.0277\]  
Thus they are positively associated in the larger population.
How are allelic associations maintained?

- selection
- non-random mating
- linkage

Again, consider two loci A and B, with alleles A₁, A₂, ..., Aₘ and B₁, B₂, ..., Bₙ occurring at frequencies p₁, p₂, ..., pₘ and q₁, q₂, ..., qₙ in the population. The haplotypes are A₁B₁, A₁B₂, ..., AₘBₙ, with frequencies ℎ₀₀, ℎ₁₁, ..., ℎₘₙ in generation 0. What is the frequency ℎᵢⱼ of haplotype AᵢBⱼ, in the next generation? To do this calculation, we make a crucial assumption of random mating in the population.

\[
h^1_{ij} = P(\text{haplotype}^1 = A_iB_j) = P(\text{haplotype}^1 = A_iB_j | \text{no recombination})P(\text{no recombination}) + P(\text{haplotype}^1 = A_iB_j | \text{recombination})P(\text{recombination}) = P(\text{haplotype}^0 = A_iB_j | \text{no recombination})(1 - \theta) + P(\text{haplotype}^0 = A_iB_j | \text{recombination})\theta = h^0_{ij}(1 - \theta) + pq_j \theta
\]

From this we can deduce that the difference in haplotype frequency between the generations is

\[
h^1_{ij} - h^0_{ij} = \theta(pq_j - h^0_{ij})
\]

When will this difference be 0?

When \(\theta = 0\) or linkage equilibrium.

We can also write this expression to characterize the difference between the true haplotype frequency and what the haplotype frequency would be under equilibrium:

\[
h^1_{ij} - pq_j = (1 - \theta)(h^0_{ij} - pq_j)
\]

Extending this to the \(k^{th}\) generation, we get:

\[
h^k_{ij} - pq_j = (1 - \theta)^k (h^0_{ij} - pq_j)
\]

We can write these as: \(D^1_{ij} = (1 - \theta)D^0_{ij}\) and \(D^k_{ij} = (1 - \theta)^k D^0_{ij}\).

The following figure shows the decline of linkage disequilibrium in a large, randomly-mating population for several different values of \(\theta\).
The rate that a randomly mating population approaches linkage equilibrium clearly depends on $\theta$. 

*Figure 4.1* Decay of linkage disequilibrium by generation.
Association studies for fine mapping

(See Sham, 4.4). Linkage analysis is a powerful tool for detecting the presence of a disease locus in a chromosomal region. However, it is not very efficient for fine mapping, since discriminating between small differences in recombination frequency requires data on a large number of informative gametes. For instance, observing no recombinant in 50 informative gametes suggests a recombination fraction of 0, but is also compatible with a recombination fraction of up to 5%. For fine mapping, linkage analysis can be followed up by association analysis. For most situations, appreciable allelic associations are only likely to exist between loci with recombination fractions of less than 1%. In the past decade or so there has been a lot of excitement about the potential power of using linkage disequilibrium to fine-map susceptibility loci.

- For older mutations, a positive association implies we are close to the susceptibility locus
  - We must have a dense marker map to get close enough to the susceptibility locus to see an association.
- If a mutation is young, there will be limited resolution since not enough time has passed to narrow the area of association

However, it is always important to remember that linkage disequilibrium can occur between loosely linked or even unlinked loci if there is a founder effect or population admixture and stratification. These possible causes of linkage disequilibrium besides tight linkage must be taken into consideration in the design, analysis, and interpretation of association studies.
Marker-based association tests in random population samples

In their simplest form, association tests are just simple applications of some generic statistical test for association between two variables:

- For dichotomous traits, this includes chi-squared tests of independence
- For continuous traits, this includes ANOVA and regression

We take a sample from the population and examine the marker alleles and the phenotype of interest on each person.

Under what conditions will this be a reasonable approach?

- trait is reasonably common
- no population sub-structure

Which of the following probabilities could be estimated from such a study?

- \( P(\text{allele}) \) (population allele frequencies)
- \( P(\text{genotype}) \) (genotype frequencies)
- \( P(D|\text{genotype}) \) (probability of disease given marker genotype)
- \( P(\text{genotype}|D) \) (probability of marker genotype given disease status)
- \( P(D) \) (disease prevalence)
Marker-based association tests in case-control samples

The most popular design for studying allelic associations between a marker and a disease is to sample based on disease status. In other words, we sample a set of cases and a set of non-diseased controls from the same population. The case-control design will be much more efficient than a random population sample for most diseases, since most diseases affect a small proportion of the population.

Which of the following probabilities could be estimated from such a study?
P(allele) (population allele frequencies)
P(genotype) (genotype frequencies)
P(D|genotype) (probability of disease given marker genotype)
P(genotype|D) (probability of marker genotype given disease status)
P(D) (disease prevalence)

We can only estimate quantities conditional on disease status when we sample based on disease status. However, this is sufficient for establishing an association.

The Idea: If a marker locus is in linkage disequilibrium with the disease locus, then the marker locus will also differ in allele frequencies between the cases and the controls.

In a simple implementation, we simply compare the distribution of marker alleles between cases and controls. This simple approach is popular for diseases with complex inheritances. Note that each individual contributes two alleles to the table.

<table>
<thead>
<tr>
<th></th>
<th>M_1</th>
<th>M_2</th>
<th>M_3</th>
<th>…</th>
<th>M_n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Controls</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

When two or more tightly linked loci are considered simultaneously for linkage disequilibrium with a complex disease, then testing for unequal allele frequencies can be replaced by testing for unequal haplotype frequencies (unequal between cases and controls). This procedure is more complicated due to indeterminacy of haplotype frequencies; we will return to it later.
Marker-based association tests in case-control samples: a simple disorder

For simple disorders, allele frequencies and penetrances can often be estimated prior to an association study, and these values can be used in a case-control design. Consider a disease locus with alleles D and N conferring high and low risk to disease with penetrances $f_{DD}$, $f_{DN}$, and $f_{NN}$. The population prevalence of the disease, assuming Hardy-Weinberg Equilibrium, is then

$$K = p_D^3 f_{DD} + 2p_D p_N f_{DN} + p_N^2 f_{NN}$$

What are the genotype frequencies among affected individuals (A)?

$$P(DD | A) = \frac{P(DD \& A)}{P(A)} = \frac{P(A | DD)P(DD)}{K} = \frac{f_{DD} p_D^2}{K}$$

Similarly,

$$P(DN | A) = \frac{P(DN \& A)}{P(A)} = \frac{P(A | DN)P(DN)}{K} = \frac{f_{DN} 2p_D p_N}{K}$$

$$P(NN | A) = \frac{P(NN \& A)}{P(A)} = \frac{P(A | NN)P(NN)}{K} = \frac{f_{NN} 2p_N^2}{K}$$

What are the genotype frequencies among unaffected individuals (U)?

$$P(DD | U) = \frac{P(DD \& U)}{P(U)} = \frac{P(U | DD)P(DD)}{1 - K} = \frac{(1 - f_{DD}) p_D^2}{1 - K}$$

$$P(DN | U) = \frac{P(DN \& U)}{P(U)} = \frac{P(U | DN)P(DN)}{1 - K} = \frac{(1 - f_{DN}) 2p_D p_N}{1 - K}$$

$$P(NN | U) = \frac{P(NN \& U)}{P(U)} = \frac{P(U | NN)P(NN)}{1 - K} = \frac{(1 - f_{NN}) p_N^2}{1 - K}$$
The above expressions quantify the differences in genotype frequencies at the disease locus between affected and unaffected individuals. From the genotype frequencies, the allele frequencies at the disease locus can then be calculated:

\[
P(D | A) = \frac{p_D^2 f_{DD} + p_D p_N f_{DN}}{K}
\]

\[
P(N | A) = \frac{p_N^2 f_{NN} + p_D p_N f_{DN}}{K}
\]

\[
P(D | U) = \frac{p_D^2 (1 - f_{DD}) + p_D p_N (1 - f_{DN})}{1 - K}
\]

\[
P(N | U) = \frac{p_N^2 (1 - f_{NN}) + p_D p_N (1 - f_{DN})}{1 - K}
\]

Now consider a marker locus B with alleles B_1, ..., B_n, with respective frequencies q_1, ..., q_n. Let the haplotype frequencies of the disease locus and the marker B be h_{D1}, ..., h_{Dn} and h_{N1}, ..., h_{Nn}. What are the conditional probabilities of being affected, given a marker genotype and assuming random mating?

\[
P(A | B_i) = \frac{f_{DD} h_{Di}^2 + f_{DN} 2 h_{Di} h_{Ni} + f_{NN} h_{Ni}^2}{q_i^2}
\]

and, for \(i \neq j\)

\[
P(A | B_j) = \frac{f_{DD} 2 h_{Di} h_{Dj} + f_{DN} (2 h_{Di} h_{Ni} + 2 h_{Dj} h_{Nj}) + f_{NN} 2 h_{Ni} h_{Nj}}{2q_i q_j}
\]

These can be thought of as the “apparent penetrances” for the marker genotypes.

We can similarly calculate the distribution of marker genotypes given disease status:

\[
P(B_i | A) = \frac{f_{DD} h_{Di}^2 + f_{DN} 2 h_{Di} h_{Ni} + f_{NN} h_{Ni}^2}{K}
\]

\[
P(B_j | A) = \frac{f_{DD} 2 h_{Di} h_{Dj} + f_{DN} (2 h_{Di} h_{Ni} + 2 h_{Dj} h_{Nj}) + f_{NN} 2 h_{Ni} h_{Nj}}{K}
\]

\[
P(B_i | U) = \frac{(1 - f_{DD}) h_{Di}^2 + (1 - f_{DN}) 2 h_{Di} h_{Ni} + (1 - f_{NN}) h_{Ni}^2}{1 - K}
\]

\[
P(B_j | U) = \frac{(1 - f_{DD}) 2 h_{Di} h_{Dj} + (1 - f_{DN}) (2 h_{Di} h_{Ni} + 2 h_{Dj} h_{Nj}) + (1 - f_{NN}) 2 h_{Ni} h_{Nj}}{1 - K}
\]

This group of expressions quantifies the differences in genotype frequencies at the marker between affected and unaffecteds. These differences can be shown to be 0 if the disease and marker loci are in linkage equilibrium, i.e. \(n_{Dj} = p_D q_i\) (Homework).

We could work out similar equations for allele frequencies at the marker as weighted sums of genotype frequencies.
These results show that differences between genotype or allele frequencies between affected and unaffected individuals in a population can arise from LD between disease and marker loci.

The data are

<table>
<thead>
<tr>
<th></th>
<th>$B_1 B_1$</th>
<th>$B_2 B_2$</th>
<th>$\ldots$</th>
<th>$B_n B_n$</th>
<th>$B_1 B_2$</th>
<th>$B_1 B_3$</th>
<th>$B_1 B_4$</th>
<th>$\ldots$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Affected</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unaffected</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For Mendelian diseases, the parameters $p_D$, $p_N q_i$, $f_{DD}$, $f_{DN}$, and $f_{NN}$ are often estimated with good precision prior to an association study. The unknown parameters in the likelihood are then only the $2n$ haplotype frequencies. The haplotype frequencies can be estimated in two ways: with and without the constraint of linkage equilibrium.

Without the constraint of linkage equilibrium, the number of parameters is $2n-2$ (because the frequencies of the haplotypes containing $D$ must sum to $p_D$, and the frequencies of the haplotypes containing $N$ must sum to $p_N$).

With the constraint of linkage equilibrium, the number of parameters is $n-1$.

The likelihood ratio test is then compared to a chi-squared distribution with $n-2$ df.
Testing for differences in the haplotype frequency distribution between cases and controls

A related, interesting question is how to test whether a specific haplotype, or any of the possible haplotypes in a gene, are associated with a phenotype? This question appears in the context of studying variants of known genes, which is a different application than disease gene mapping.

Scientists are interested in haplotypes because sometimes the effect of mutations (e.g. SNPs) depend strongly on whether the mutations occur on the same chromosome (in cis position, as a haplotype) or on opposite homologous chromosomes (in trans position). In other words, the haplotype may be highly associated with disease even if the individual markers are not.

If we are interested in whether the haplotype distribution differs between cases and controls, then we have

\[ H_0: \text{ haplotype frequency distribution is equal } (p_1, \ldots, p_m) \]
\[ H_1: \text{ haplotype frequency distribution differs between cases and controls. We need one distribution } (q_1, \ldots, q_m) \text{ for cases and another distribution } (r_1, \ldots, r_m) \text{ for controls.} \]

However, we only get data on genotypes, not haplotypes. Consider data from two loci.

We can use a likelihood ratio test to compare these hypotheses:

\[ LRTS = -2 \ln \left( \frac{L_{null}(p_1, \ldots, p_m)}{L_{cases}(q_1, \ldots, q_m)L_{controls}(r_1, \ldots, r_m)} \right) \]

As usual, since the null hypothesis is nested within the alternative hypothesis the LRTS is asymptotically chi-squared. How many degrees of freedom will the test have?

\[ m - 1 \]

The added difficulty is estimating the haplotype frequencies. This difficulty exists whether we consider the case and control populations separately, or combined.
Consider data from 1000 individuals on two biallelic loci:

<table>
<thead>
<tr>
<th>Genotype at Locus 1</th>
<th>11</th>
<th>12</th>
<th>22</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>10</td>
<td>110</td>
<td>271</td>
</tr>
<tr>
<td>12</td>
<td>57</td>
<td>309</td>
<td>102</td>
</tr>
<tr>
<td>22</td>
<td>76</td>
<td>55</td>
<td>10</td>
</tr>
</tbody>
</table>

Genotype at Locus 2

Which haplotypes can be discerned with certainty, where does ambiguity arise?

Ambiguity arises in the middle cell, the double heterozygote.
If we think of the haplotypes in the middle cell as “missing” data, we see this is a natural application of the EM algorithm.

Start with some initial estimates of the haplotype frequencies. For example, assume each haplotype is equally likely.

1. From the haplotype frequencies, find the expected number of haplotypes in the “middle” cell. (Expectation Step)
2. Use the imputed “complete data,” where all haplotype counts are known, to estimate the haplotype frequencies. (Maximization Step)

EM applied to the data on the previous page (thanks to Stephanie Monks):

<table>
<thead>
<tr>
<th>Iteration</th>
<th>p12</th>
<th>p21</th>
<th>p11</th>
<th>p22</th>
<th>Expected # with phase 11/22</th>
<th>Expected # with phase 12/21</th>
<th>Sum of Squared Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>154.5</td>
<td>154.5</td>
<td>0.250000</td>
</tr>
<tr>
<td>1</td>
<td>0.45</td>
<td>0.21</td>
<td>0.17</td>
<td>0.17</td>
<td>71</td>
<td>238</td>
<td>0.056757</td>
</tr>
<tr>
<td>2</td>
<td>0.50</td>
<td>0.25</td>
<td>0.13</td>
<td>0.12</td>
<td>35</td>
<td>274</td>
<td>0.006990</td>
</tr>
<tr>
<td>3</td>
<td>0.51</td>
<td>0.27</td>
<td>0.11</td>
<td>0.11</td>
<td>24</td>
<td>285</td>
<td>0.001278</td>
</tr>
<tr>
<td>4</td>
<td>0.52</td>
<td>0.27</td>
<td>0.11</td>
<td>0.10</td>
<td>22</td>
<td>287</td>
<td>0.000118</td>
</tr>
<tr>
<td>5</td>
<td>0.52</td>
<td>0.28</td>
<td>0.10</td>
<td>0.10</td>
<td>21</td>
<td>288</td>
<td>0.000008</td>
</tr>
<tr>
<td>6</td>
<td>0.52</td>
<td>0.28</td>
<td>0.10</td>
<td>0.10</td>
<td>21</td>
<td>288</td>
<td>0.000000</td>
</tr>
</tbody>
</table>

The procedure was terminated when the estimates of the haplotype frequencies barely changed between iterations:

\[ \sum_{j=1}^{4} (\hat{p}_j^{\text{iter}} - \hat{p}_j^{\text{iter-1}})^2 < 0.000001. \]
Empirical P-values/Permutation P-values

When the overall sample size is small and/or the haplotype probabilities are small, p-values based on the asymptotic distribution can be inaccurate. Because of this, it is desirable to have another way to get a p-value for the LRTS rather than comparing it to a Chi-square distribution.

Recall that a p-value is the probability of observing something as extreme or more extreme than the observed test statistic if the null hypothesis is true.

What is our null hypothesis?

What permutations of the data correspond to this null hypothesis?

Permutation test for a data set with $N_{\text{cases}}$ and $N_{\text{controls}}$:

1. Compute the test statistic on the observed data. This yielded the observed test statistic: $T^{\text{obs}}$
2. For $i=1,\ldots,R$ where $R$ is a large number, e.g. 10,000
   a. Randomly assign "case" status to $N_{\text{cases}}$ individuals from the full sample. The remaining individuals are "controls."
   b. Compute the statistic on the permuted data, $T^{\text{perm}}_i$
3. The p-value is estimated to be
   $$\text{empirical p-value} = \frac{\# T^{\text{perm}}_i \geq T^{\text{obs}}}{R}$$

An important choice in any permutation test is what to use as the test statistic. Different test statistics can be more or less sensitive to the kinds of differences one is interested in. For more information about permutation tests for comparing haplotype distributions, see:

Zhao JH, Curtis D, Sham PC (2000) Model-free analysis and permutation tests for allelic associations. Hum Heredity 50:133-139

There is, however, a limitation to combining permutation-testing with likelihood based results. It is very computationally expensive to estimate three likelihoods for cases, controls, and combined for a larger number of different permutations of the data.
Other Haplotype Tests of Association

The test of association we just discussed is based on a comparison of the haplotype frequency distributions. If the test is significant, this indicates that the distribution of haplotypes differs between cases and controls.

However:
- The test provides no information on the associations between the individual haplotypes and the trait.
- The parameteric LRTS test statistic can be unreliable, and permutation-test p-values can be too computationally expensive to be feasible.

Another tool we can use to relate haplotypes to a dichotomous trait is logistic regression. Schaid et al.² proposed using logistic regression along with a score test for assessing significance. Advantages of this approach are:
- Along with a “global” test of association, we can compute score test statistics for individual haplotypes.
- Score statistics are rapid to compute since they do not rely on maximizing a likelihood. This makes permutation-based p-values practical.

Suppose we have a dichotomous trait and would like to use logistic regression to relate the trait with the number of copies of each haplotype:

\[
\log\text{it}(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{m-1} x_{m-1}
\]

Here, \(x_i\) is the number of haplotypes of type \(i\) for an individual, \(x_i = 0, 1, \text{ or } 2\). \(p\) is the probability of being a case.

What does the logistic regression reduce to under the null hypothesis of no association between the disease and the haplotypes?

\[
\log\text{it}(p) = \beta_0
\]

What are the null and alternative hypotheses, in terms of the \(\beta\) coefficients, for a test of no association versus association?

\[H_0: \beta_1 = \beta_2 = \ldots = \beta_{m-1} = 0\]
\[HA: \text{ Some (at least one) } \beta_i \text{ not equal to 0}\]

What is the interpretation for \(\beta\)?

The odds of being a disease increase multiplicatively by a factor of \(e^{\beta_i}\) for each additional copy of the \(i^{th}\) haplotype.

² Schaid DJ et al. (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 70: 425-434
Example: Schaid et al. (2002) examined the association between HLA alleles and immune response following vaccination for measles. (For biological reasons, the HLA alleles are logical genes to study for their possible effect on immune response.) 220 unrelated people were evaluated for antibody levels. This outcome can be treated quantitatively, or dichotomized at a threshold considered to be a clinically negative response.

Three (of 11) HLA loci showed significant association with the binary outcome variable. The loci are DQB, DRB and HLA-B and have 12, 11 and 24 alleles, respectively. A total of 678 haplotypes were enumerated that were consistent with the observed data; however, only 178 had non-zero frequency estimates. Pooling rare haplotypes (freq < 0.005) into one group resulted in 40 haplotypes to evaluate.

Schaid et al. (2002) give the following table describing the haplotypes most strongly associated with both binary and quantitative traits. Parametric and permutation-based p-values were quite similar in this case.

**Table 2**

<table>
<thead>
<tr>
<th>HAPLOTYPES</th>
<th>BINARY TRAIT</th>
<th>QUANTITATIVE TRAIT</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAPLOTYPES</td>
<td>SCORE</td>
<td>$\chi^2$</td>
</tr>
<tr>
<td>HAPLOTYPES</td>
<td>SCORE</td>
<td>$\chi^2$</td>
</tr>
<tr>
<td>DQB</td>
<td>DRB</td>
<td>HLA-B</td>
</tr>
<tr>
<td>63</td>
<td>13</td>
<td>60</td>
</tr>
<tr>
<td>31</td>
<td>4</td>
<td>44</td>
</tr>
<tr>
<td>21</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>21</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>21</td>
<td>3</td>
<td>8</td>
</tr>
</tbody>
</table>