The loss of statistical power to distinguish populations when certain samples are ambiguous

Martin O'Hely and Montgomery Slatkin*

Department of Integrative Biology, University of California, 3060 Valley Life Sciences, Bldg. 3140 (4151-4155 VLSB), Berkeley, CA 94720-3140, USA

Received 28 August 2002

Abstract

Case–control studies are used to map loci associated with a genetic disease. The usual case–control study tests for significant differences in frequencies of alleles at marker loci. In this paper, we consider the problem of comparing two or more marker loci simultaneously and testing for significant differences in haplotype rather than allele frequencies. We consider two situations. In the first, genotypes at marker loci are resolved into haplotypes by making use of biochemical methods or by genotyping family members. In the second, genotypes at marker loci are not resolved into haplotypes, but, by assuming random mating, haplotypes can be inferred using a likelihood method such as the expectation–maximization (EM) algorithm. We assume that a causative locus has two alleles with a multiplicative effect on the penetrance of a disease, with one allele increasing the penetrance by a factor \( p \).

We find, for small values of \( p / 1 \) and large sample sizes, asymptotic results that predict the statistical power of a test for significant differences in haplotype frequencies between cases and a random sample of the population, both when haplotypes can be resolved and when haplotypes have to be inferred. The increase in power when haplotypes can be resolved can be expressed as a ratio \( R \), which is the increase in sample size needed to achieve the same power when haplotypes are resolved over when they are not resolved. In general, \( R \) depends on the pattern of linkage disequilibrium between the causative allele and the marker haplotypes but is independent of the frequency of the causative allele and, to a first approximation, is independent of \( p \).

For the special situation of two di-allelic marker loci, we obtain a simple expression for \( R \) and its upper bound.

Keywords: Haplotypes; Case–control; Likelihood ratio test; Gene mapping; Complex disease

1. Introduction

In a case–control study of a disease, individuals are classified into two groups: cases (individuals with the disease) and controls (individuals without the disease). Genotypes of marker loci are obtained from members of each group. Marker alleles that are significantly more frequent in cases than in controls are probably closely linked to causative alleles. Population stratification, meaning the presence of subgroups of individuals in the sample population who have different frequencies of marker alleles and different disease prevalences, can complicate the analysis (Clayton, 2001) but new methods can correct for stratification if enough marker loci can be surveyed (e.g. Pritchard et al., 2000; Devlin et al., 2001).

Haplotype rather than allele frequencies can be compared (Chapman and Wijsman, 1998; Valdes et al., 1999; Kaplan and Morris, 2001a, b). Although greater statistical power can be achieved using haplotype frequencies (Chapman and Wijsman, 1998), resolving the genotypes of individuals who are heterozygous at two or more loci into component haplotypes is difficult unless parental genotypes are available also. Biochemical methods for obtaining haplotypes are currently so expensive and time consuming that they cannot be used on large samples.

Several methods exist for inferring haplotypes from multilocus genotypes in a randomly mating population. Hill (1974) developed a maximum likelihood (ML) method and introduced an algorithm, later called the expectation–maximization (EM) algorithm (Dempster et al., 1977), for inferring haplotype frequencies. Hill (1974) found that, for two di-allelic loci and large sample sizes, his ML method achieved no better than double the accuracy as in the situation in which
haplotypes are resolved. Hill (1975) showed how this method could be generalized to more loci and more alleles. Recently, several groups have implemented the EM algorithm and tested its performance (Excoffier and Slatkin, 1995; Kidd et al., 1998). Stephens et al. (2001) have introduced a Bayesian method in which the prior distribution is generated by a coalescent simulation and have shown that it performs better than the EM algorithm when large numbers of loci are surveyed.

In this paper, we consider the performance of ML estimation of haplotype frequencies in a case–control study. The question is how much statistical power is lost in a case–control study if haplotypes cannot be resolved, or, equivalently, how much do sample sizes need to be increased when haplotypes cannot be resolved, to achieve the same power as when haplotypes can be resolved. We can obtain general analytic results for large sample sizes when the causative allele has a multiplicative effect on the probability of having the disease. Large sample sizes are needed to use asymptotic methods leading to the non-central chi-square distribution. Multiplicative effects are needed to ensure that haplotypes in the cases are randomly associated if they are in the base population.

The comparisons we will be making may be thought of as similar to those made by Service et al. (1999), who develop a test which includes haplotype information and find it compares favourably with a method due to Terwilliger (1995) which does not consider haplotype data. Terwilliger’s method does, however, incorporate some knowledge of the underlying haplotype structure (specifically knowledge of the map distances between the markers), which can be thought of analogously to the structural information needed for the EM algorithm (specifically which haplotype pairings are ambiguous).

The advantage of our method for comparing power with and without resolved haplotypes is that we use two methods based on the same likelihood framework, and so obtain explicit formulae allowing comparisons of the two situations, allowing us to explore the comparison under a variety of evolutionary assumptions without the need of simulations.

Section 2 establishes some notation and terminology, and details an example that we hope will be of use in framing the general theory. Section 3 presents the main results for arbitrary numbers of marker loci. Section 4 then applies the general results when there are two or three di-allelic marker loci. Section 5 considers bounds for the increase in power with two marker loci when haplotypes can be resolved. Sections 6 and 7 present proofs of results stated in Sections 3 and 5, respectively. Section 8 discusses results and relates them to other theoretical studies. The Appendix gives a brief description of the general theory we have used in treating the asymptotic distribution of log-likelihood ratio (LLR) statistics.

2. Terminology and notation

Suppose we are studying a genetic disease and we are interested in finding a locus whose alleles affect an individual’s susceptibility to the disease. Marker alleles which are physically linked to this causative locus might be expected to show an association with the disease, and conversely markers which show association with the disease might be suspected to be physically linked to the causative locus. Such an association might be indicated by differences in marker frequencies between cases and controls for the disease. Suppose that we consider diallelic markers two at a time, denoting the alleles at one locus as $B$ and $b$, and at the other $C$ and $c$. We can view these two markers as a single marker with four alleles $BC$, $Bc$, $bC$, $bc$; but there arises the complication that, in diploid individuals, it is difficult to distinguish individuals carrying $BC$ and $bc$ “meta-alleles,” or haplotypes, from those carrying $Bc$ and $bC$. Without direct information about this phase resolution, it is possible to make statistical inferences of the haplotype frequencies in each of the case and control populations, but the statistical power this provides to determine if a significant difference between haplotype frequencies exists will be lower than in the situation where the resolution is available. We consider the question: how much lower?

The problem described above can be generalized: consider two populations, each of which can be categorized into $k$ classes, where sampling from the population can only be accomplished by drawing a random group of $d$ members at a time. In the situation of the previous paragraph, the populations are the chromosomes of individuals with the disease, and of the population as a whole (strictly this would be a case–random study, not case–control), the classes are the four haplotypes, and with diploids we must, by definition, take two chromosomes at a time. Further suppose that there exist sets of combinations of the classes which, when drawn in the sample, are indistinguishable, either because they are truly indistinguishable or because for, say, economic or technical reasons it is not possible to make the distinction. In our example, the combination of $BC$ and $bc$ is indistinguishable from that of $Bc$ and $bC$. We wish to test the hypothesis that the distributions are different against a null that the distributions are identical, using a log-likelihood ratio (LLR) test. The question of interest is, how much statistical power is lost because of the inability to distinguish the precise distribution of types in a particular sample? A related and ultimately equivalent question is, how much does the power change depending on one’s choice of which combinations to resolve?

We will investigate the situation where the difference between the compositions of the populations is small: we will quantify the difference by a single parameter, and effectively compute the power of testing with and
The distinction is that the final element listed in the haplotypes. Each haplotype will fall into one of the classes, denoted \( F \) in the example where \( k = 4 \) and the set of classes is \( \{ BC, Bc, bC, bc \} \). The sample, comprising a \( d \)-element multi-set of haplotypes, will be called a genotype—we use a multi-set since the multiplicity of haplotypes in a sample is important but there is no sense of ordering of the haplotypes in the sample. \( \{ BC, Bc \} \) and \( \{ BC, bc \} \) are examples of genotypes; \( d \) in general will represent the ploidy and should probably be thought of as just 2. A set of genotypes which are indistinguishable will be called a phenotype; \( \{ \{ BC, bc \}, \{ Bc, bC \} \} \), as well as singletons containing each of the remaining eight genotypes, are the phenotypes in the example. Each genotype is a member of precisely one phenotype.

Denote a phenotype by \( \phi \), and write \( \Phi \) for the set of all phenotypes under a particular structure of indistinguishability among the genotypes. Write \( g \) for the cardinality of \( \Phi \). Write \( \Phi' \) for the set of phenotypes, under the structure where all genotypes are uniquely identifiable, and denote the cardinality of \( \Phi' \) by \( g' \). For any genotype \( \phi' \), write \( \mathcal{G}(\phi') \) for the set of genotypes which cannot be distinguished from \( \phi' \) under the structure \( \Phi \).

Under our motivating example,

\[
\Phi = \{ \{ BC, BC \}, \{ BC, Bc \}, \{ BC, bC \}, \{ Bc, Bc \}, \{ bC, Bc \}, \{ Bc, bC \} \}, \\
\Phi' = \{ \{ BC, BC \}, \{ BC, Bc \}, \{ BC, bC \}, \{ Bc, Bc \}, \{ bC, Bc \}, \{ Bc, bC \} \}.
\]

(The distinction is that the final element listed in \( \Phi \) is a two-element set.)

For any phenotype \( \phi \), define a \( \binom{d}{0} \)-tensor \( M_\phi \) by requiring that its action on the vectors \( (e_1, \ldots, e_d) \) result in a 1 if \( i_1, \ldots, i_d \) are in \( \phi \) and 0 otherwise (if \( e_i \) is a \( k \)-vector consisting of all zeroes except for a 1 in the \( i \)-th position). Since the phenotypes are multi-sets, not ordered \( d \)-tuples, \( M_\phi \) will be invariant under permutation of its arguments. Write \( M_\phi^p \) for the one-form which maps a vector \( v \) to \( M_\phi(p, \ldots, p, v) \).

In terms of the example, and respectively with the description of \( \Phi \) above, the tensors can be represented by \( 4 \times 4 \) matrices (the tensors’ actions on column vectors \( v_1 \) and \( v_2 \) are given by \( v_1^T M v_2 \)):

\[
\begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0
\end{bmatrix}
\begin{bmatrix}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0
\end{bmatrix}
\begin{bmatrix}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0
\end{bmatrix}
\]

In \( \Phi' \) the list is the same except the final tensor is split into

\[
\begin{bmatrix}
0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0
\end{bmatrix}
\]
The distribution of haplotypes within a population will be described by a \( k \)-dimensional vector \( \mathbf{p} = (p_1, ..., p_k) \) where \( p_i \) is the frequency of haplotype \( i \). (Note that since these frequencies add to unity, one might only specify \( (p_1, ..., p_{k-1}) \).) We will assume that drawing any particular genotype is equivalent to independently drawing its constituent haplotypes, so that the frequency of any phenotype \( \phi \) will be given by \( M_\phi(p, ..., p) \) and written \( p_\phi \) for brevity. The result of taking \( m \) samples from this population will be described by the vector \( (\mu_{\phi_1}, ..., \mu_{\phi_g}) \) where \( \mu_{\phi_i} \) is the number of samples which fall into the phenotype \( \phi_i \).

For a single population, the sampling density is
\[
f((\mu_{\phi_1}, ..., \mu_{\phi_g}); \mathbf{p}) = \left( \begin{array}{c} m \\ \mu_{\phi_1}, ..., \mu_{\phi_g} \end{array} \right) \prod_{i=1}^g p_{\phi_i}^{\mu_{\phi_i}},
\]
and with two populations, with haplotype distributions \( \mathbf{p} \) and \( \mathbf{q} \), taking \( m \) and \( n \) samples from the populations respectively, the sampling distribution is
\[
f((\mu_{\phi_1}, ..., \mu_{\phi_g}); (\mathbf{v}_{\phi_1}, ..., \mathbf{v}_{\phi_g}); \mathbf{p}, \mathbf{q}) = \left( \begin{array}{c} m \\ \mu_{\phi_1}, ..., \mu_{\phi_g} \end{array} \right) \left( \begin{array}{c} n \\ \mathbf{v}_{\phi_1}, ..., \mathbf{v}_{\phi_g} \end{array} \right) \prod_{i=1}^g p_{\phi_i}^{\mu_{\phi_i}} q_{\phi_i}^{\mathbf{v}_{\phi_i}},
\]
where \( (\mathbf{v}_{\phi_1}, ..., \mathbf{v}_{\phi_g}) \) are the counts of phenotypes observed in the samples from the \( \mathbf{q} \) population.

Consider taking \( N \) such samples from the two-population system (so that in all \( mN \) phenotypes are obtained from the \( \mathbf{p} \) population and \( nN \) from the \( \mathbf{q} \) population—we specify this awkward sampling method so as to provide \( N \) independent identically distributed samples, as required for the likelihood ratio test, while allowing the \( \mathbf{p} \) and \( \mathbf{q} \) populations to have sizes in the ratio \( m:n \)). Determine maximum-likelihood estimates for the parameters \( \mathbf{p} \) and \( \mathbf{q} \), firstly under the assumption that they are equal, and secondly without this assumption. (When haplotypes are directly available, this is just a matter of evaluating the likelihood functions using the sample haplotype frequencies, while the EM algorithm is useful when haplotypes must be inferred.) Form the quotient of the corresponding likelihoods and consider twice the logarithm of this quotient. This quantity is the log-likelihood ratio statistic. This is known to be distributed as a non-central chi-squared random variable with \( k \) degrees of freedom (Wald, 1943), asymptotically for large \( N \). The formula for the non-centrality parameter is given in the appendix; it may be useful to the reader to think of the non-centrality parameter as the value the log-likelihood ratio statistic attains when the data are set equal to their expectations under the alternative as in Morris and Kaplan (2002).

Once the non-centrality parameter has been determined, the power of the test can be computed by determining the probability the non-central chi-squared distribution assigns to values greater than the critical value of the test, i.e. the value above which the (central) chi-squared distribution assigns a probability equal to the error rate desired. We will not explicitly compute the power; we will restrict ourselves to questions of when the two varieties of tests have equivalent power. Since the tests will have equal degrees of freedom, this will come down to equating the non-centrality parameters.

3. Results

Suppose that the distribution of haplotypes in the \( \mathbf{q} \) population is a perturbation of the distribution in the \( \mathbf{p} \) population, that is
\[
\mathbf{q} = \mathbf{p} + \Delta(x),
\]
where \( \Delta : \mathbb{R} \rightarrow \mathbb{R}^k \) is continuous and differentiable in a neighbourhood of 0, and \( \Delta(0) = 0 \). Write \( y = (y_1, ..., y_k) \) for the vector of derivatives with respect to \( x \) of \( \Delta(x) \).

Theorem 1a. Under an ambiguity structure \( \Phi \), the log-likelihood ratio statistic will be asymptotically distributed as a non-central chi-squared random variable with \( k - 1 \) degrees of freedom and non-centrality parameter
\[
\frac{mnNz^2d^2}{m+n}r
\]
to second order in \( z \), where
\[
r = \sum_{\phi \neq \phi} \frac{1}{p_{\phi}} (M_{\phi}^\mathbf{p}y)^2.
\]

Theorem 1b. Under the ambiguity structure \( \Phi' \), where each phenotype corresponds to a unique genotype, the log-likelihood ratio statistic will be asymptotically distributed as a non-central chi-squared random variable with \( k - 1 \) degrees of freedom and non-centrality parameter
\[
\frac{mnNz^2d^2}{m+n}r'
\]
to second order in \( z \), where
\[
r' = \sum_{\phi \neq \phi} \frac{1}{p_{\phi}'} (M_{\phi}^\mathbf{p}y)^2 = \frac{1}{d} \sum_{i=1}^k \frac{y_i^2}{p_i}.
\]

Note that the expression for the approximation of the non-centrality parameter shows how the parameter depends separately on sampling choices \( (m, n, N) \), the degree of difference between the distributions \( (z) \), the structure of haplotype frequencies in the two populations \( (r, \mathbf{p}, \mathbf{q}, \Delta) \), and the sampling structure \( (d, M_h) \).

The main content of Theorem 1b is the very last equality, since otherwise it is a special case of Theorem 1a. The value of the alternate form of \( r' \) is evident in considering the numbers of terms in the sums; they are the numbers of phenotypes and haplotypes. In the
Corollary 2. The difference $r' - r$ is equal to

$$\sum_{\phi' \in \mathcal{G}(\phi)} \left( \sum_{\phi'' \in \mathcal{G}(\phi')} \frac{1}{p_{\phi''}} (M_{\phi''}^p y)^2 - \frac{1}{p_{\phi'}} (M_{\phi'}^p y)^2 \right)$$

$$= \sum_{\phi' \in \mathcal{G}(\phi)} \left( \sum_{\phi'' \in \mathcal{G}(\phi') \setminus \mathcal{G}(\phi)} \frac{(p_{\phi''} (M_{\phi''}^p y) - p_{\phi'} (M_{\phi'}^p y))^2}{p_{\phi'} p_{\phi''} p_{\phi'}} \right).$$

This effectively quantifies the difference in the power between the situations where ambiguities are resolved and when they are not. It will ultimately be interesting to consider the ratio between $r$ and $r'$, or equivalently, $(r' - r)/r'$.

The Corollary follows from Theorems 1a and b in a straightforward way: the first line comes from performing the subtraction and noting that single-genotype phenotypes cancel out. The second line makes use of the observations that $M_{\phi}^p = \sum_{\phi' \in \mathcal{G}(\phi)} M_{\phi'}^p$ and $p_{\phi} = \sum_{\phi' \in \mathcal{G}(\phi)} p_{\phi'}$, and the identity

$$\sum_{i=1}^{n} y_i \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} y_i} = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \left( y_i y_j - \frac{y_i y_j}{\sum_{k=i}^{n} y_k} \right).$$

4. An application to complex disease mapping

This section will be concerned with applying the results described in Section 3 to our “motivating example”. We have given an informal description of underlying marker haplotype structure but have not yet discussed its relation with the causative allele or the relation of this allele to disease status. Determining results for two markers is straightforward, and we will present results for three markers so as to indicate what the result will be for a general number of markers.

4.1. The model

We assume that an autosomal locus with two alleles, $A$ and $a$, affects the probability that an individual has a particular genetic disease (or some other binary trait). The causative allele, $A$, has a multiplicative effect on penetrance. Individuals with genotype $AA$ have a probability of having the disorder increased by a factor of $\pi$, and individuals with genotype $Aa$ have a probability of having the disorder increased by a factor of $\pi^2$, both relative to individuals with genotype $aa$. For $A$ to be truly causative we should have $\pi > 1$ but our analytic theory will not require that restriction.

In addition to the $A/a$ locus we will consider two di-allelic marker loci $B/b$ and $C/c$ that have no effect on penetrance. We do not specify any gene order. The state of the population is determined by the eight haplotype frequencies $p_{ABC}, \ldots, p_{abc}$. We assume random mating, so that genotype frequencies in the base population are the products of the appropriate haplotype frequencies.

Although the haplotype frequencies completely characterize the population, it will be more convenient to use other parameters instead. We will use the three allele frequencies $(p_A, p_B, p_C)$, the extent of linkage disequilibrium (LD) between $B$ and $C$ ($D_{BC}$), and three measures of the extent of LD between $A$ and the marker haplotypes ($D_{ABC}, D_{A|BC}, D_{A|BC}$). The relationships among these variables are shown in Table 1. The reason for this choice is that it separates quantities describing the marker loci, $p_B, p_C$, and $D_{BC}$, which are observable, from those involving the causative locus, which are non-observable and are regarded as parameters. We will consider the implications for our results of various values of $p_A$ but not in general assume knowledge of the LD coefficients between the causative allele and the marker haplotypes.

We will later have occasion to describe the LD between the causative allele and the marker haplotypes in terms of its magnitude (this may be thought of as $(D_{ABC}^2 + D_{ABC}^2 + D_{ABC}^2 + D_{ABC}^2)^{1/2}$) and direction (a description of the relative ratios of these LD coefficients). A geometric interpretation of these notions is available: the vector $(D_{ABC}, D_{A|BC}, D_{A|BC}, D_{A|BC})$ can be thought of as being in $\mathbb{R}^4$, and the requirement that $D_{ABC} + D_{A|BC} + D_{A|BC} + D_{A|BC} = 0$ shows this vector is in fact in the three-dimensional subspace where the sum of the co-ordinates is zero. Then the magnitude and direction of the LD are just the magnitude and direction of this vector in three space.

We assume two sample sets are available, a random sample from the base population (the randoms), and a sample of individuals with the disease (the cases). We use randoms instead of controls (individuals who do not have the disease) because the calculations are substantially easier. For a rare disease, there is little difference between randoms and controls, and particularly for a late-onset disease, establishing that unaffected individuals will not develop the disease is problematic.

Clearly among the randoms the marker haplotype frequencies are

$$p_{BC} = p_{ABC} + p_{aBC},$$
$$p_{Bc} = p_{Abc} + p_{abC},$$
$$p_{bc} = p_{Aabc} + p_{abC},$$
$$p_{bc} = p_{Aabc} + p_{abC},$$
Table 1  
Translation between LD and haplotype frequency parameterizations

<table>
<thead>
<tr>
<th>Disequilibrium parameters</th>
<th>Haplotype frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_A = p_{ABC} + p_{ABC} + p_{ABC} + p_{ABC} )</td>
<td>( p_{ABC} = p_A p_B p_C - p_A D_{BC} + D_{ABC} )</td>
</tr>
<tr>
<td>( p_B = p_{ABC} + p_{ABC} + p_{ABC} + p_{ABC} )</td>
<td>( p_{ABC} = p_A p_B (1-p_C) + p_A D_{BC} + D_{ABC} )</td>
</tr>
<tr>
<td>( p_C = p_{ABC} + p_{ABC} + p_{ABC} + p_{ABC} )</td>
<td>( p_{ABC} = p_A (1-p_B) p_C + p_A D_{BC} + D_{ABC} )</td>
</tr>
<tr>
<td>( D_{ABC} = p_{ABC} - p_A (p_{ABC} + p_{ABC}) )</td>
<td>( p_{ABC} = p_A (1-p_B)(1-p_C) - p_A D_{BC} + D_{ABC} )</td>
</tr>
<tr>
<td>( D_{Bc} = p_{Bc} - p_A (p_{Bc} + p_{Bc}) )</td>
<td>( p_{ABC} = (1-p_A) p_{ABC} - (1-p_A) D_{BC} + D_{ABC} )</td>
</tr>
<tr>
<td>( D_{Abc} = p_{Abc} - p_A (p_{Abc} + p_{Abc}) )</td>
<td>( p_{ABC} = (1-p_A)(1-p_B) p_C + (1-p_A) D_{BC} - D_{ABC} )</td>
</tr>
<tr>
<td>( D_{abc} = p_{abc} - p_A (p_{abc} + p_{abc}) )</td>
<td>( p_{ABC} = (1-p_A)(1-p_B)(1-p_C) - (1-p_A) D_{BC} - D_{ABC} )</td>
</tr>
</tbody>
</table>

while some calculation reveals that among the cases the corresponding frequencies are

\[
q_{BC} = \frac{p_{ABC} + p_{ABC}}{p_A + p_A}, \\
q_{BC} = \frac{p_{ABC} + p_{ABC}}{p_A + p_A}, \\
q_{BC} = \frac{p_{ABC} + p_{Abc}}{p_A + p_A}, \\
q_{bc} = \frac{p_{Abc} + p_{Abc}}{p_A + p_A},
\]

where \( p_A = p_{ABC} + p_{ABC} + p_{ABC} + p_{ABC} \) and \( p_a = 1 - p_A \) are the population frequencies of alleles at the susceptibility locus, cf. Table 1.

4.2. Applying Theorem 1 and Corollary 2 to this situation

Here \( \pi - 1 \) plays the role of \( \alpha \). We have

\[
\frac{\partial (q - p)}{\partial \pi} \bigg|_{\pi = 1} = \begin{bmatrix} D_{A|BC} \\ D_{A|Bc} \\ D_{A|bc} \end{bmatrix} = \begin{bmatrix} D_{A|BC} \\ D_{A|Bc} \\ D_{A|bc} \end{bmatrix}
\]

and so \( y = (D_{A|BC}, D_{A|Bc}, D_{A|bc})^T \). We consider an ambiguity structure with \( \{BC, bc\} \) indistinguishable from \( \{Bc, bc\} \), and discover

\[
r' = \frac{1}{2} \left( \frac{D_{A|BC}}{p_{BC}} + \frac{D_{A|Bc}}{p_{Bc}} + \frac{D_{A|bc}}{p_{bc}} \right),
\]

\[
r = r' - \frac{p_{BC} p_{Bc} p_{bc}}{2(p_{BC} p_{Bc} + p_{BC} p_{bc})} \left( \frac{D_{A|BC}}{p_{BC}} + \frac{D_{A|Bc}}{p_{Bc}} - \frac{D_{A|bc}}{p_{bc}} \right) \cdot (D_{A|BC} - D_{A|Bc} - D_{A|bc})^2.
\]

4.3. Extending to three markers

Suppose now there is a third marker locus with alleles \( D \) and \( d \). The population will contain (up to) eight haplotypes and the ambiguity structure will account for six pairs of double heterozygotes and the six pairs taken from the four triple heterozygotes. Then
The structure in $r'$ should be clear: for each marker haplotype there is a term whose denominator is the frequency of the haplotype and whose numerator is the square of the LD between the haplotype and the causative allele. In $r$ the haplotypes in the numerator and the squared term correspond to those which may be confused under the ambiguity structure (with the particular pairing corresponding to the signs in the squared term). The denominators are the total frequencies of the phenotypes containing the ambiguous pairs. The first six terms represent the six double heterozygotes, whose phenotypes each contain two genotypes, while the remaining six terms represent the triple heterozygote phenotype, which contains four genotypes, and the squared term corresponds to those which may causative allele. In $r$ the square of the LD between the haplotype and the marker haplotypes in the numerator and whose numerator is the frequency of the haplotype and whose denominator is the squared term corresponds to those which may be confused under the ambiguity structure (with the particular pairing corresponding to the signs in the squared term). The denominators are the total frequencies of the phenotypes containing the ambiguous pairs. The first six terms represent the six double heterozygotes, whose phenotypes each contain two genotypes, while the remaining six terms represent the triple heterozygote phenotype, which contains four genotypes, from which there are $\binom{6}{2} = 6$ distinct pairs of haplotypes.

For four diallelic markers, $r'$ will have 16 terms while $r' - r$ will involve 24 terms arising from double heterozygotes, 48 from triple heterozygotes, and 28 from the quadruple heterozygotes. It is possible to automate the generation of the formulae for an arbitrary number of diallelic loci and evaluate the ratio $r' / r$ with simulated data. Fig. 2 presents the results of doing just this: for 2, ..., 7 diallelic marker loci we randomly generated one million sets of multi-locus haplotype frequencies and evaluated $R$ at each, generating the cumulative distribution functions shown. While the precise shapes of these distributions must be interpreted with caution (since parameters were generated essentially uniformly and not according to any biological model) it appears plausible that $R$ is rarely more than 3 with three, four or five marker loci and likewise rarely more than 4 for six or seven marker loci. It also seems plausible to adopt as a rule of thumb that with two loci, in a good number of cases there may be no advantage at all to directly determining haplotypes; with three loci there is always an advantage, although it may be arbitrarily small; while with four or more loci there is always an advantage corresponding to an $R$ of at least one plus a small amount, 1.1, 1.4 and 1.75 (for 4, 5, 6 and 7 loci, respectively). Fig. 2 also shows the cdf for the case of two triallelic loci; qualitatively it resembles the three diallelic loci case for least advantage to haplotyping but shares the two diallelic loci upper limit of $R < 2$.

5. Two markers: the power gained by resolving haplotypes

In this section we return to the case of two markers. We want to compare the power resulting from non-centrality parameters $r'$ and $r$. It will be convenient to consider the ratio $R = r' / r$ since the distribution of the LLR statistic in the cases of haplotypes resolved and not resolved differs only in the non-centrality parameter, and so (asymptotically) equal power will be available from the two methods if, when $N_h$ sets of $m+n$ individuals are sampled and have their haplotypes resolved, or $N_n$ sets of $m+n$ individuals are sampled and their haplotypes are not resolved, $N_h r' = N_n r$. Thus for equal power when not resolving haplotypes, $R$ times as many individuals need be sampled.

Since both $r'$ and $r' - r$ are pure quadratics in the $D_{ij}[r]$, the ratio $R$ does not depend on the magnitude of LD between the causative allele and the marker haplotypes, but only on the direction of this LD. Furthermore, the formulae for $r'$ and $r$ are effectively independent of the causative allele frequency $p_A$, with the equivocation necessary since $p_A$ does affect the quantities by restricting the range of LD allowable between the causative locus and marker haplotypes. However, since the sign of the LD cancels in the computation of $R$, we conclude that $R$ is independent of both this LD magnitude and $p_A$. A weak effect may
remain due to the potential for \( p_A \) to be correlated with the direction of LD.

One can compare the accuracy of the approximation of the ratio between the haplotyping and genotyping situations to the same ratio with the exact values from Section 6.1. The two are plotted for two distinct haplotype frequency distributions in Fig. 1. In the situation of Fig. 1a, the exact and approximate ratios remain within half a percent of each other over the range of \( \pi \) shown, while in the situation of Fig. 1b, the ratios remain within 10% of each other (data not shown).

We can show

**Theorem 3.** For fixed marker haplotype frequencies \( R \) is maximized when

\[
D_{A|BC} = D_{A|bc} = - D_{A|bc} = - D_{A|BC}
\]

for which

\[
R = \left( 1 - p_{BC}p_B p_{bc}p_C (p_{BC}^{-1} + p_B^{-1} + p_{bc}^{-1} + p_C^{-1}) \right)^{-1}
\]

For minor marker alleles \( B \) and \( C \) with positive linkage disequilibrium, \( R < 2 \). Regardless of the linkage disequilibrium between \( B \) and \( C \), the maximum value of \( R \) is the greater of \( (p_{BC} + p_{bc})^{-1} \) and \( (p_{BC} + p_{bc})^{-1} \).

The result for positive LD shows that the resolution of haplotypes provides no greater power than doubling the size of the samples. The condition in Eq. (4) merits some interpretation. The equality of the first and third terms imply \( D_{A|B} \), the disequilibrium between \( A \) and \( B \), is zero. Likewise equality of the first and last terms implies \( D_{A|C} \) is zero. Thus, haplotyping will be most advantageous over genotyping in a situation where testing for associations with the individual marker alleles separately has no power whatsoever. This would presumably be a highly pathological situation.

### 6. Proof of Theorems 1

This proof will be in several stages. First, we will apply the method outlined in the appendix to obtain the exact asymptotic distribution. This will be done for a general pair of populations described by haplotype frequency vectors \( p \) and \( q \). Then we will determine the second-order approximation to the non-centrality parameter by expanding its exact expression in a Taylor series. Finally, we will demonstrate the rewriting of \( r' \) from Theorem 1b.

#### 6.1. The asymptotic distribution

Our situation may be described in the terms used in the appendix by taking \( \Theta = (p_1, \ldots, p_{k-1}, q_1, \ldots, q_{k-1}) \) as the parameters of the consolidated population (leaving \( p_k = 1 - p_1 - \cdots - p_{k-1} \) and similarly \( q_k \)). The equations describing the null hypothesis are \( p_i - q_i = 0 \), \( i = 1, \ldots, k - 1 \), and so

\[
\mathcal{E}(\Theta) = [I - I]^T,
\]

\[
\rho(\Theta) = p - q.
\]

From the appendix we have that the distribution will be a non-central chi-squared with \( k - 1 \) degrees of freedom, and the real work of this section will be the computation of the non-centrality parameter.

From the sampling distribution (1) the matrix \( C(\Theta) \) has as its \((i,j)\)-element

\[
E\left( \frac{\partial^2 \log f}{\partial \theta_i \partial \theta_j} \right) = - \sum_{\phi \in \Phi} \left( E(\mu_{\phi_i}) \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p_{\phi_i} + E(\mu_{\phi_j}) \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log q_{\phi_j} \right),
\]

Clearly \( E(\mu_{\phi_i}) = mp_{\phi_i} \) while

\[
\frac{\partial}{\partial p_i} p_{\phi_i} = \frac{\partial}{\partial p_i} M_{\phi_i}(p, \ldots, p) = dM_{\phi_i} \left( \frac{\partial p_i}{\partial p_i}, p, \ldots, p \right) = dM_{\phi_i} (e_i - e_k, p, \ldots, p)
\]

(where the symmetry of \( M_{\phi} \) has been exploited, and \( e_i \) is the canonical basis vector with zeroes for all but the \( i \)th entries, and unity for the \( i \)th) and

\[
\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p_{\phi_i} = \left( \frac{\partial}{\partial \theta_i} \frac{\partial p_{\phi_i}}{p_{\phi_i} \partial \theta_j} + \frac{\partial^2 p_{\phi_i}}{p_{\phi_i} \partial \theta_i \partial \theta_j} \right) = - \frac{1}{p_{\phi_i}} \frac{\partial p_{\phi_i}}{\partial \theta_i} \frac{\partial^2 p_{\phi_i}}{\partial \theta_i \partial \theta_j} + \frac{1}{p_{\phi_i}} \frac{\partial^2 p_{\phi_i}}{\partial \theta_i \partial \theta_j}
\]

so that

\[
E\left( \frac{\partial^2 \log f}{\partial \theta_i \partial \theta_j} \right) = - \sum_{\phi \in \Phi} \left( \frac{m}{p_{\phi_i}} \frac{\partial p_{\phi_i}}{\partial \theta_i} \frac{\partial q_{\phi_i}}{\partial \theta_j} + \frac{n}{q_{\phi_i}} \frac{\partial q_{\phi_i}}{\partial \theta_i} \frac{\partial q_{\phi_i}}{\partial \theta_j} \right) + \frac{m}{p_{\phi_i}} \frac{\partial^2 p_{\phi_i}}{\partial \theta_i \partial \theta_j} + \frac{n}{q_{\phi_i}} \frac{\partial^2 q_{\phi_i}}{\partial \theta_i \partial \theta_j}
\]

since

\[
\sum_{\phi \in \Phi} p_{\phi} = \sum_{\phi \in \Phi} q_{\phi} = 1.
\]

Note that since the \( p_{\phi} \) depend only on the \( p_i \) among the \( \theta_i \), and similarly the \( q_{\phi} \) only on the \( q_i \), \( C(\Theta) \) is block diagonal:

\[
C(\Theta) = \begin{bmatrix} mP & 0 \\ 0 & nQ \end{bmatrix},
\]
where $P$ has elements
\begin{equation}
\sum_{\phi \in \Phi} \left( \frac{1}{p_\phi} \frac{\partial p_\phi}{\partial \theta^i} \frac{\partial p_\phi}{\partial \theta^j} \right)
\end{equation}
for $i, j = 1, \ldots, g - 1$ and similarly for $Q$ (with $i, j = g, \ldots, 2g - 2$). As a consequence, the non-centrality parameter $N \rho(\Theta)^T (\Xi(\Theta)^T C(\Theta)^{-1} \Xi(\Theta))^{-1} \rho(\Theta)$ has the structure
\begin{equation}
N \rho^T \left( \frac{1}{m} P^{-1} + \frac{1}{n} Q^{-1} \right)^{-1} \rho.
\end{equation}

It is possible to describe the matrix $P$ (and, analogously, $Q$) by a tensor-vector formula, rather than specifying its elements as in (5). Write $D_\phi$ for the column vector with entries
\[
\frac{\partial p_\phi}{\partial \theta^i} = \Delta M_\phi (e_i - e_k, p, \ldots, p), \quad i = 1, \ldots, k - 1.
\]

Then
\[
\frac{\partial p_\phi}{\partial \theta^i} \frac{\partial p_\phi}{\partial \theta^j}
\]
are the elements of the matrix $D_\phi D_\phi^T$, and so
\[
P = \sum_{\phi \in \Phi} \frac{1}{p_\phi} D_\phi D_\phi^T.
\]

Recall $M_\phi^p$ is the one-form mapping $v \rightarrow M_\phi^p (v, p, \ldots, p)$. We will abuse notation by writing $M_\phi^p$ for the row-vector representation of the one-form as well, so that the action of $M_\phi^p$ on $v$ is just the inner product $M_\phi^p v$. Then
\[
D_\phi = d(M_\phi^p)^T,
\]
where
\[
J = \begin{bmatrix}
1 & 0 & \cdots & 0 \\
0 & 1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 1 \\
-1 & -1 & \cdots & -1
\end{bmatrix},
\]
and so
\begin{equation}
P = d^2 \sum_{\phi \in \Phi} \frac{1}{p_\phi} J^T (M_\phi^p)^T M_\phi^p J.
\end{equation}

6.2. Second-order approximation

Expand the non-centrality parameter in a Taylor series around $\alpha = 0$: using $X$ in place of $(P^{-1}/m + Q^{-1}/n)^{-1}$ in (6) for ease of writing
\[
\frac{\partial}{\partial \alpha} (\rho^T X \rho) = \frac{\partial \rho^T}{\partial \alpha} X \rho + \rho^T \frac{\partial X}{\partial \alpha} \rho + \rho^T X \frac{\partial \rho}{\partial \alpha} = 0
\]
(since $\rho = 0$ when $\alpha = 0$) and
\[
\frac{\partial^2}{\partial \alpha^2} (\rho^T X \rho) = \frac{\partial^2 \rho^T}{\partial \alpha^2} X \rho + \rho^T \frac{\partial^2 X}{\partial \alpha^2} \rho + \rho^T X \frac{\partial^2 \rho}{\partial \alpha^2} X + 2 \rho^T \frac{\partial X}{\partial \alpha} \frac{\partial \rho}{\partial \alpha} X + 2 \rho^T X \frac{\partial X}{\partial \alpha} \frac{\partial \rho}{\partial \alpha}
\]
and, again since $\rho = 0$ when $\alpha = 0$, all terms except the second last one are zero. Moreover, $P = Q$ at $\alpha = 0$ so
\[
\frac{\partial^2}{\partial \alpha^2} \left( \rho^T \left( \frac{1}{m} P^{-1} + \frac{1}{n} Q^{-1} \right)^{-1} \rho \right) = \left. \frac{2 m n \rho^T}{m + n} \frac{\partial \rho}{\partial \alpha} \right|_{\alpha = 0} \cdot \frac{\partial \rho}{\partial \alpha}.
\]
Thus, for small $\alpha$, the non-centrality parameter is approximately
\[
\frac{m n \alpha^2 \rho^T}{m + n} \left. \frac{\partial \rho}{\partial \alpha} \right|_{\alpha = 0} \cdot \frac{\partial \rho}{\partial \alpha} = \frac{m n \alpha^2 \rho^T}{m + n} \theta^0. \quad \text{and thus the non-centrality parameter for the test will be}
\]
\[
\frac{m n \alpha^2 \rho^T}{m + n} \sum_{\phi \in \Phi} \frac{1}{p_\phi} (M_\phi^p \theta^0)^2.
\]

6.3. Re-writing $r'$

Recall that this is the situation where all genotypes resolve to unique haplotypes. The phenotypes correspond to ordered partitions $n_1 + \cdots + n_k = d$ of $d$ into a sum of $k$ non-negative integers, representing the number of each haplotype present among the $d$ in the sample. For each such phenotype, the corresponding tensor $M_\phi^p$ has $d!/(n_1! \cdots n_k!)$ non-zero “entries”. Thus, for this phenotype,
\begin{equation}
p_\phi = \begin{pmatrix}
d \\
n_1, \ldots, n_k
\end{pmatrix} p_{n_1}^{p_1} \cdots p_{n_k}^{p_k}
\end{equation}
while $M_\phi^p \theta^0$ is obtained by expanding out (8) into a sum of $d!/(n_1! \cdots n_k!)$ terms, each of which is the product of $d$
The remaining summation is
\[ \sum_{n_1 + \ldots + n_k = d-1} n_i \left( \begin{array}{c} d-1 \\ n_1, \ldots, n_k \end{array} \right) p_1^{n_1} \ldots p_i^{n_i} \ldots p_k^{n_k} \]
\[ = \sum_{n_1 + \ldots + n_k = d-1} n_i \left( \begin{array}{c} d-1 \\ n_1, \ldots, n_k \end{array} \right) \times \prod_{n_i = 1}^{n_i \geq 1} \frac{(d-2)!}{n_i! \ldots (n_i+1)! \ldots n_k!} \times \prod_{n_i = 1}^{n_i \geq 1} p_i^{n_i} \]
\[ = (d-1) p_i \sum_{n_1 + \ldots + n_k = d-2} \frac{(d-2)!}{n_1 \ldots n_k} p_1^{n_1} \ldots p_i^{n_i} \ldots p_k^{n_k} \]
\[ = (d-1) p_i (p_1 + \ldots + p_k)^{d-2} \]
and so the coefficient of \( y_i^2 \) will be
\[ \frac{1}{p_i d} (1 + p_i (d-1)) \]

The cross term \( y_i y_j \) will have coefficient
\[ \sum_{n_1, n_j \geq 1} \frac{2}{n_1 + \ldots + n_k = d} \left( \begin{array}{c} d-1 \\ n_1, \ldots, n_i - 1, \ldots, n_k \end{array} \right) \times \frac{p_1 \ldots p_k}{p_i p_j} \]
\[ = 2 \sum_{n_1, n_j \geq 1} \frac{d-1}{n_1 \ldots n_j} \frac{d-1}{d-1} \times \left( \begin{array}{c} d-2 \\ n_1, \ldots, n_i - 1, \ldots, n_j - 1, \ldots, n_k \end{array} \right) \times \frac{p_1^{n_1} \ldots p_i^{n_i-1} \ldots p_j^{n_j-1} \ldots p_k^{n_k}}{p_i^{n_i} \ldots p_j^{n_j} \ldots p_k^{n_k}} \]
\[ = 2 \frac{d-1}{d} (p_1 + \ldots + p_k)^{d-2} \]
\[ = 2 \frac{d-1}{d} \]

Summing the \( y_i y_j \) terms across all pairs of \( i \) and \( j \) gives
\[ \frac{d-1}{d} ((y_1 + \ldots + y_k)^2 - y_1^2 - \ldots - y_k^2) \]
\[ = \frac{d-1}{d} (y_1^2 + \ldots + y_k^2) \]
since the \( y_i \) add to zero.

Combining the results of the last two paragraphs reveals that the overall expression for \( r' \) can be written without cross terms as
\[ \frac{1}{d} \left( \frac{y_1^2}{p_1} + \ldots + \frac{y_k^2}{p_k} \right) \]

6.4. Re-writing \( r' \) when \( d = 2 \)

The proof of the result in the previous section is complicated and it may be of interest to see the computation when \( d = 2 \).

Here the \( M_{iy}^x \) are of two types: those consisting of homozygotes (if \( \phi \) is the homozygous \( i \) genotype then \( M_{iy}^x = p_i y_j \)) and those consisting of heterozygotes (if \( \phi \) is the heterozygote containing \( i \) and \( j \) then \( M_{iy}^x = p_j y_j + p_j y_i \)). Thus, the expression for \( r' \) gives
\[ \sum_{i=1}^{k} \frac{(p_i y_i)^2}{p_i^2} + \sum_{i=1}^{k} \sum_{j=i+1}^{k} \frac{(p_j y_j + p_j y_i)^2}{2p_i p_j} \]
\[ = \sum_{i=1}^{k} y_i^2 + \sum_{i=1}^{k} \sum_{j=i+1}^{k} \frac{p_j y_j^2 + p_j y_i^2}{2p_i} + y_i y_j \]
\begin{equation}
\sum_{i=1}^{k} y_i^2 + \left( \sum_{i=1}^{k} \sum_{j=1}^{k} p_j y_{ij} - \frac{1}{2} \sum_{i=1}^{k} y_i^2 \right)
+ \sum_{i=1}^{k} \sum_{j=1}^{k} y_j y_{ij}
= \frac{1}{2} \left( \sum_{i=1}^{k} y_i \right)^2 + \sum_{i=1}^{k} \sum_{j=1}^{k} p_j y_{ij}^2
= 0 + \frac{1}{2} \left( \sum_{i=1}^{k} p_i \right) \left( \sum_{j=1}^{k} y_j^2 / p_j \right)
= \frac{1}{2} \left( \sum_{j=1}^{k} y_j^2 / p_j \right),
\end{equation}

7. Proof of Theorem 3

Maximizing \( r'/r \) is equivalent to maximizing \( (r' - r)/r' \), which given Eqs. (2) and (3) is a much more amenable expression:

\begin{equation}
r' - r
= \frac{p_{BC} p_{Bc} + p_{Bc} p_{bc}}{p_{BC} p_{Bc} + p_{Bc} p_{bc}} \left( \frac{D_{ABC} + D_{A|BC} - D_{A|BC} + D_{A|bc}}{p_{BC} + p_{Bc} + p_{bc}} \right)^2
\end{equation}

with equality if and only if \( D_{A|BC} = D_{A|bc} = -D_{A|Bc} = -D_{A|Bc} \).

The inequality is equivalent to

\begin{equation}
\left( \frac{D_{A|BC} + D_{A|bc} - D_{A|BC} + D_{A|bc}}{p_{BC} + p_{Bc} + p_{bc}} \right)^2
\leq \left( \frac{D_{A|BC}^2 + D_{A|bc}^2 + D_{A|BC}^2 + D_{A|bc}^2}{p_{BC} + p_{Bc} + p_{bc}} \right)
\times \left( \frac{1}{p_{BC}} + \frac{1}{p_{Bc}} + \frac{1}{p_{bc}} \right)
\end{equation}

which in turn is equivalent to

\begin{equation}
\frac{2D_{A|BC}D_{A|bc}}{p_{BC} p_{bc}} + \frac{2D_{A|BC}D_{A|bc}}{p_{Bc} p_{bc}} + \frac{2D_{A|BC}D_{A|bc}}{p_{BC} p_{bc}}
- \frac{2D_{A|BC}D_{A|BC}}{p_{BC} p_{bc}} - \frac{2D_{A|BC}D_{A|bc}}{p_{Bc} p_{bc}} - \frac{2D_{A|bc}D_{A|bc}}{p_{BC} p_{bc}}
\leq \frac{D_{A|BC}^2 + D_{A|bc}^2}{p_{BC} p_{bc}} + \frac{D_{A|BC}^2 + D_{A|bc}^2}{p_{Bc} p_{bc}} + \frac{D_{A|BC}^2 + D_{A|bc}^2}{p_{BC} p_{bc}}
\end{equation}

and this is readily seen to be equivalent to

\begin{equation}
\frac{(D_{A|BC} - D_{A|bc})^2}{p_{BC} p_{bc}} + \frac{(D_{A|BC} - D_{A|bc})^2}{p_{Bc} p_{bc}}
+ \frac{(D_{A|BC} + D_{A|bc})^2}{p_{BC} p_{bc}} + \frac{(D_{A|BC} + D_{A|bc})^2}{p_{Bc} p_{bc}}
+ \frac{(D_{A|BC} + D_{A|bc})^2}{p_{BC} p_{bc}} + \frac{(D_{A|BC} + D_{A|bc})^2}{p_{Bc} p_{bc}} \geq 0.
\end{equation}

The claim follows since (10) is patently true. It is also clear that equality will occur under the stated conditions. Note that the term on the left of (10) can also be written

\begin{equation}
\frac{(D_{A|BC} - D_{A|BC})^2}{p_{BC} p_{bc}} + \frac{(D_{A|BC} - D_{A|BC})^2}{p_{Bc} p_{bc}} + \frac{2D_{A|BC}}{p_{BC} p_{bc}} + \frac{2D_{A|BC}}{p_{Bc} p_{bc}} + \frac{2D_{A|BC}}{p_{BC} p_{bc}}
\end{equation}

so it is clear that equality occurs for \( D_{A|BC} = D_{A|BC} = 0. \)

7.2. Maximization with positive \( D_{BC} \)

The term on the right of (9) can itself be analysed for a maximum. For simplicity, rewrite the term as

\begin{equation}
\frac{abc + abd + acd + bcd}{ad + bc}
\end{equation}

with \( a + b + c + d = 1 \). Clearly, this will be non-negative (and equal to zero in the event \( a = b = 0 \)), and since it is equal to

\begin{equation}
\frac{ad(b + c) + bc(a + d)}{ad + bc}
\end{equation}

it will be less than unity (equaling unity when one of the variables is itself equal to one and the others are zero). This, of course, is a situation which would never occur in practice, since it would require one of the marker alleles to have frequency zero.

Attempting to maximize the expression using multivariate calculus will, of course, be futile (the extrema have already been determined), but it will afford some insight. Setting \( a = 1 - b - c - d \) and solving for the partial derivatives equal to zero, one finds stationary points for \( b + d = 1/2 \) and \( c = d \), as well as \( c + d = 1/2 \) and \( b = d \). These result in a value of 1/2 for the expression. To verify that this is indeed a maximum (for
the moment setting aside the fact, established in the last paragraph, that it isn’t) one would attempt to prove
\[
\frac{ad(a + d) + bc(b + c)}{ad + bc} \geq \frac{1}{2}
\]
which is equivalent to
\[
ad(a + d) + bc(b + c) \geq \frac{1}{2} ad + \frac{1}{2} bc
\]
which is in turn equivalent to
\[
ad(a + d - \frac{1}{2}) \geq bc(\frac{1}{2} - (b + c)).
\]
Now \(a + d - 1/2 = 1/2 - b - c\) since \(a + b + c + d = 1\).
Supposing, without loss of generality, the expression is unchanged by exchanging \(a\) for \(b\) and \(c\) for \(d\), that \(a + d \geq 1/2\), an equivalent inequality is then
\[
ad \geq bc.
\]
Naïvely this last ineqeuation seems reasonable—after all, it was assumed that of the pairs \(a, d\) and \(b, c\) the former is “larger.” However (and just as well in the light of what was shown above) one could set \(a\) to any value greater than a half and let \(d\) decrease to zero while keeping \(b = c\) to falsify (12). Nevertheless, provided (12) holds (where it is assumed \(a + d > b + c\)), one has that \(r’/r \leq 2\).

One circumstance where the conditions in the previous paragraph hold is when there is positive disequilibrium between the minor alleles at the marker loci. In such a situation, \(p_{BC} + p_{bc} - p_{Bc} - p_{bC} = (p_B - p_b)(p_C - p_c) + 4D_{BC}\) which is clearly positive (minor alleles and positive LD), and \(p_{BC}p_{bc} - p_{Bc}p_{bC} = D_{BC}(p_B - p_b)(p_C - p_c)\) is also clearly positive.

Notice that a slightly negative \(D_{BC}\) will falsify (12), but not \(a + d > b + c\). It is natural to ask whether sufficiently negative \(D_{BC}\) will reverse the latter inequality and so restore the criterion for \(R < 2\). Recall (Lewontin, 1964) that \(D_{BC} > -p_{BPC}\). For any negative \(D_{BC}\), \(p_{BC}p_{bc} - p_{Bc}p_{bC}\) will be negative and so we investigate whether \(p_{BC} + p_{bc} - p_{Bc} - p_{bC}\) can be negative for sufficiently negative \(D_{BC}\). This would require
\[
4p_{BC}p_{bc} > (p_B - p_b)(p_C - p_c),
\]
which is equivalent to \(p_B + p_C > 1/2\). Thus, when \(p_B + p_C > 1/2\), the LD between \(B\) and \(C\) can be sufficiently negative to again force \(R < 2\). This corresponds with the intuitive payload of \(p_{BC} + p_{bc} - p_{Bc} - p_{bC} = (p_B - p_b)(p_C - p_c) + 4D_{BC}\), which is that for common marker alleles (so a smaller difference between the major and minor alleles), not very negative \(D_{BC}\) is required to make the quantity negative.

7.3. Maximization in general

Writing \(f(a, b, c, d)\) for the expression in (11) and continuing to assume that \(a + d > b + c\), it is possible to show that \(f(a, b, c, d) \leq f(a + d, (b + c)/2, (b + c)/2, 0) = a + d\). Firstly, \(f(a, b, c, d) \leq f(a, (b + c)/2, (b + c)/2, d)\); this is equivalent to
\[
\left(\frac{ad + (b + c)^2}{2}(a + d)\right)(ad + bc)
\]
and in turn
\[
\left(\frac{1}{2}\right) ad(a + d - (b + c)) \geq abcd(a + d - (b + c)),
\]
under the assumption \(a + d > b + c\) this is equivalent to the arithmetic–geometric means inequality for \(b\) and \(c\) (and thus equality occurs for \(b = c\)). Secondly, \(f(a, b, b, d) \leq f(a + d, b, b, 0)\); this is equivalent to \(ab^2 + 2abd + b^2d \leq (a + d)(ad + b^2)\) and in turn
\[
2b \leq a + d
\]
which is true by assumption. Here, equality occurs for \(b + c = a + d\), when the limit is again found to be 2.

8. Discussion

Our results are summarized by two expressions, \(r (3)\) and \(r’ (2)\), that indicate the relative statistical power in a case–control study when haplotypes can be resolved and when they must be inferred using ML. These expressions assume that each causative allele increases the probability of having a disease by a factor \(\pi\); they are correct to order \((\pi - 1)^2\). Our numerical analysis shows that the analytic approximation is accurate over a range of values of \(\pi\) that is biologically important. If \(\pi\) is very large, it is likely that a causative allele with such major effect would have been already detected in studies of affected families or in case–control studies in which one marker locus at a time is analysed. The ratio \(R = r’/r\) indicates the proportional gain in statistical power when haplotypes are resolved instead of inferred. \(R\) can be regarded as the increase in sample size necessary when haplotypes cannot be resolved to achieve the same power as when haplotypes can be resolved. While \(R\) is an approximation, it has the desirable property that it is independent from both the frequency of the causative allele and the magnitude of the LD between the causative allele and the marker haplotypes, both of which would be unknown in the context of gene mapping.

With more than 2 million SNP markers now available in the human genome (Sachidanandam et al., 2001), subsets of those markers can be chosen to increase the chances of success in a case–control study. Assuming equal cost of genotyping each marker locus, the choice of marker loci depends on the frequencies of the minor
alleles and on the extent of LD between them. If haplotypes can be resolved, then the marker loci should be chosen so that \( r' \) is as large as possible. For two marker loci, the expression for \( r' \) can be rewritten as a sum of terms of the form

\[
P_A^2 \sum_h \frac{(p_{h/A} - p_h)^2}{p_h},
\]

where \( \sum_h \) represents the sum over \( h \in \{ BC, Bc; bC, bc \} \) and \( p_{h/A} = p_{Ah}/p_A \) is the frequency of marker haplotype \( h \) among chromosomes carrying the susceptibility locus. The generalization to more than two marker loci is obvious. Similarly,

\[
r' - r = \left( \frac{1}{p_{BC}p_{bc}} + \frac{1}{p_{BC}p_{bc}} \right)^{-1}
\times P_A^2 \left( \frac{p_{BCA}}{p_{BC}} + \frac{p_{BcA}}{p_{Bc}} - \frac{p_{BcA}}{p_{bc}} \right)^2.
\]

Both \( p_A \) and the \( p_{h/A} \) are of course unknown. If \( p_A \) is small, meaning that the causative allele is rare, \( P_A^2 \) is very small and it will be difficult to detect significant differences in haplotype frequencies unless one or more of the \( p_{h/A} - p_h \) is large enough to offset \( P_A^2 \). The division by \( p_h \) suggests that \( p_{h/A} \gg p_h \) will be more helpful than \( p_{h/A} \ll p_h \) in providing statistical power. Such a large increase in the frequency of a particular haplotype on \( A \)-bearing chromosomes is possible only if that haplotype is rare in the population. If \( A \) is rare, it is likely to have arisen recently and still retain ancestral haplotypes at closely linked loci, so it is reasonable to assume that there are such marker loci present, although finding them requires considerable luck. For rare \( A \), it is unlikely that marker haplotypes that are relatively common will be of much use. The potential increase in frequency of a common haplotype on \( A \)-bearing chromosomes is too small to compensate for the small value of \( P_A^2 \).

If \( A \) is more common, then implications of our results are different. It is preferable to choose markers that are relatively common and among which there is little or no linkage disequilibrium. That would ensure that all of the marker haplotype frequencies are similar and would give roughly equal change of any of the terms in the sum to be large. If \( A \) is common, then rare markers would appear to be of little use. While their presence in the denominator terms would be useful, for common \( A \) and rare \( h \) the terms \( p_{h/A} \) and \( p_h \) will both be small, and once squared, would cancel out the power provided by the division. Thus more common marker haplotypes are to be preferred. However, having one rare marker haplotype will not actually cause us to lose power; in fact, it would enable the remaining three haplotypes to be more frequent.

Conditions for a large \( r' - r \) should be clear. The \( P_A^2 \) term can be ignored since it is a scaling factor present in the expression for \( r' \) also. The term with the reciprocals shows that the difference will be large only if neither class of doubly heterozygous individuals in the general population is small, which seems reasonable since otherwise double heterozygotes should be accurately inferred. The term with the conditionals in it suggests that the ratios of the \( p_{h/A} \) to the \( p_h \) now matter more than the differences. To make \( r' - r \) large, a standout \( p_{h/A} \gg p_h \) is required, together with the absence of such a standout among the haplotypes in the other double heterozygote. Basically, if the power in \( r' \) is coming from a single \( p_{h/A} \gg p_h \), then resolving haplotypes gives a more substantial boost in power. If it is coming from a pair of haplotypes, then if those haplotypes pair to give a double heterozygote, resolving helps, while if the opposite is true, resolving will be of little help. Lastly if power in \( r' \) is due to a \( p_{h/A} \ll p_h \) then resolving the haplotypes will be of little use.

Likewise we can consider conditions for a small, and possibly zero \( r' - r \). The reciprocal term will make \( r' - r \) small when one of the marker haplotypes is vanishingly rare, as discussed in the previous paragraph. The term involving conditional frequencies will be zero for

\[
P_A ABC - P_{A|BC} = P_{A|Bc} - P_{A|bc}
\]

(note the reversal of the conditioning, \( P_{A|h} = p_{Ah}/p_h \)); that is, the presence of the \( C \) marker gives the same marginal change in information about \( A \) regardless of the state of the \( B/b \) locus, and vice-versa. Thus although \( C \) does carry some information on the causative locus, it does not add to that which is carried by \( B \).

Whether causative alleles tend to be rare or common is a question that, in the absence of adequate data, can be debated only on theoretical grounds (e.g. Pritchard, 2001; Reich and Lander, 2001). Our point here is that if haplotypes can be completely resolved, the optimal choice of markers in a case–control study depends on the unknown frequency of the causative allele. If causative alleles tend to be rare, then successful case–control studies will require very good luck because the right low-frequency haplotypes will have to be used. If causative alleles tend to be common, less luck is needed and relatively common marker haplotypes should be used.

More generally, our results are conditioned on the distribution of marker haplotype frequencies and their association with the causative allele, and particularly the direction of the LD between the causative allele and the markers. In practice this cannot be known, so it will be necessary to determine the distribution of haplotype frequencies conditional on readily selectable marker characteristics (frequency and recombination fraction, for example). This would require theoretical or simulation modelling of the evolutionary processes underlying the population genetics of the loci in question, or the analysis of an empirical haplotype data set. The
parameters for such models, or the scope of real data, have the potential to change drastically depending on the circumstances of interest (e.g., between fine mapping and genome-wide screens). Because of the complications involved, we have not addressed these questions beyond the limited considerations in the previous paragraphs.

When marker haplotypes cannot be completely resolved, as is currently true in virtually all case-control studies, our results can be summarized by \( R = r'/r \), which indicates the increase in sample size needed to achieve the same statistical power as when haplotypes can be resolved. The expression for \( R \) provides relatively simple guidelines for understanding the effects of choosing markers of various types. If marker loci are chosen so that one marker haplotype is rare, then \( R \) is nearly 1 indicating that little power is lost because of the inability to resolve haplotypes. The reason is that, when one haplotype is rare or missing, there is almost complete LD at the marker loci and an ML method such as the EM algorithm is most successful in that situation. If marker loci are chosen so that all marker haplotypes are roughly equally common, then LD among the marker loci is small and usually \( R < 2 \) when considering two marker loci. Therefore, in the two situations of most practical interest, little statistical power is lost when marker haplotypes have to be inferred.

There has been recent work analysing issues similar to those we consider. Kaplan and Morris (2001a, b), consider the power available from pairs of diallelic markers in two situations: firstly comparing genotype data at each locus separately and using Bonferroni correction to construct a single test out of the two locus-wise tests, and secondly combining information from the two loci to obtain, essentially, a four-allele marker locus. This latter case corresponds to our motivating example in the event of fully resolved haplotypes. With simulated data, they found “no clear advantage” using such meta-loci. In fact, their results generally indicate somewhat less power with meta-loci than by combining single-locus tests. They assume an additive model for the inheritance of the disease, that is, the probability an Aa heterozygote contracts the disease is arithmetically intermediate to the similar probabilities for the two homozygotes. Our multiplicative model requires geometric intermediacy.

Akey et al. (2001) consider case-control analysis using single diallelic markers as well as haplotypes across two and four such markers, without Bonferroni correction. They find an increase in power as more marker loci are used in the test. They assume an absence of phenocopies, that is (in our notation) individuals homozygous for the a allele at the causative locus have zero probability of contracting the disease, and present results only for recessive and dominant modes of inheritance, although their theory allows for intermediate dominance. It is not clear whether their observed increase in power would survive correction for multiple testing.

Xiong et al. (2002) present a generalized \( T^2 \)-test which is designed to detect differences in the joint distributions of single-locus genotypes in individuals across several loci, or in chromosomes across loci, and make use of it to determine extended haplotypes which are associated with disease status. The data in the two situations correspond to our haplotypes-not resolved versus—resolved situations, and multiplicative penetrance and case-random design are implicitly assumed. They find no difference in power when haplotypes can be resolved—the power of the two tests is mathematically equal. Closer analysis of their results reveals that their haplotypes-resolved test statistic depends only on the genotype information. They also provide a single-marker example in which their test provides increased power over the conventional goodness of fit \( \chi^2 \)-test; our likelihood ratio test can be shown to give equal power to the Xiong et al. in this case. For two and more markers, the Xiong et al. method may or may not provide greater statistical power than the method we present for haplotypes not resolved, depending on the particular set of marker and causative allele haplotypes underlying the system. For clarity, we will consider the two marker locus case. The reason there is no clear “winner” between our method and theirs is that they are testing a different null to ours. Their null is described as an absence of LD between the causative allele and each of the markers separately; in the notation we have used, this can be seen to be equivalent to \( p_B - q_B = p_C - q_C = 0 \). With only two constraints, their method requires two degrees of freedom compared to our method’s three, and in certain circumstances this provides their method an advantage. However, their null is a superset of ours, and so other circumstances would permit our method to retain power while theirs has none. This situation is \( D_{A|BC} = D_{A|bc} = -D_{A|Bc} = -D_{A|bc} \) (or equivalently \( D_{A|B} = D_{A|C} = 0 \), cf. Section 7.1 where this was shown to be the condition among the LDs between causative allele and marker haplotype which gave maximal advantage to directly resolved haplotypes), where one might say that the LD between the causative allele and the marker haplotypes is “strongly haplotypic”. We note that our method can be adapted to consider the same null hypothesis as Xiong et al.; in this case our method has equal power to theirs. Furthermore, in this case our method, like theirs, does not give any increase in power when haplotypes are resolved.

Chapman and Wijmjan (1998) consider multi-allelic versus di-allelic markers. (Our haplotypes-resolved situation is equivalent to a four-allele marker.) They assume a specific evolutionary model and establish that multiple alleles always provide a more powerful test for case-control studies, as well as providing evidence that
between six and eight alleles is the point of diminishing returns from increasing the allele count. They also show that better power is available with equifrequent alleles, give an example where the addition of two rare alleles in fact decreases the power, and show that the improvement in power from using more markers is greater when the genetic association between disease and markers is weaker.

It is interesting that there is such a diversity of conclusions drawn from the above studies as to the utility of haplotype methods. Kaplan and Morris (2001a, b) suggest even fully resolved haplotypes give no advantage over single marker methods, while Akey et al. (2001) show an advantage. This discrepancy may be explainable simply by considering whether appropriate type I error rates have been used. Xiong et al. (2002) provide a method that combines information from multiple marker loci directly and is expected to provide in general greater power than ad hoc combination of these data, but which is unable to take any advantage of haplotype information. Chapman and Wijsman (1998) illustrate clearly that, while greater numbers of alleles may well increase the amount of information that is present in the data, they can require higher degrees of freedom tests which can lower the power available. Ultimately, there is a need for information on whether the true state of causative and marker alleles is likely to benefit more from methods whose power comes from restricting the degrees of freedom, or from refining the null hypothesis used to model the absence of an association.

Although all of the above studies use non-central chi-squared distributions in their power calculations, all but Xiong et al. (2002) use these in the context of goodness-of-fit tests. (Xiong et al., utilize the non-central chi-squared as the distribution of the \( T^2 \) statistic under their alternative hypothesis.) Our study generates the non-central chi-square as the distribution of the LLR for gene mapping. We thank Chad Garner, Laurent Excoffier for helpful comments about this project, and three anonymous referees for their contributions to the preparation of the manuscript.

**Appendix. Asymptotic distribution of the LLR statistic**

The determination of the distribution of the LLR statistic for large numbers of observations was considered by Wald (1943): suppose a stochastic system depends on the \( k \) parameters \( \Theta = (\theta^0, \ldots, \theta^k) \), and one wishes to test for rejection of a null hypothesis described by the equations \( \xi^0(\Theta) = \cdots = \xi^r(\Theta) = 0 \ (r < k) \). Make \( N \) independent observations of a random variable \( x \) whose distribution has density \( f(x, \Theta) \). Given these observations, determine the maximum likelihoods with either \( \Theta \) free \( (\ell_1) \) or constrained by the null hypothesis \( (\ell_0) \), and form the log likelihood ratio statistic \( \log(l_1/l_0) \). Provided certain technical conditions are met (essentially the maximum likelihood estimates of the parameters should asymptotically follow a multivariate normal distribution), then asymptotically as the number of observations \( N \) becomes large, this statistic will be distributed as a non-central chi-squared random variable with \( r \) degrees of freedom. The non-centrality parameter will be (the lone element of the \( 1 \times 1 \) matrix)

\[
N \rho(\Theta)^T \Xi(\Theta) \Xi^{-1}(\Theta)^{-1} \rho(\Theta),
\]

where \( C, \rho \) and \( \Xi \) are \( k \times k, \ r \times 1 \) and \( k \times r \) matrices respectively, described by

\[
C(\Theta)_{ij} = E \left( - \frac{\partial^2 \log(f(x, \Theta))}{\partial \Theta^i \partial \Theta^j} \right),
\]

\[
\Xi(\Theta)_{r} = \frac{\partial \xi^r}{\partial \Theta}, \quad \rho(\Theta)_{1} = \xi^0(\Theta).
\]

Under the null hypothesis, the non-centrality will be zero (since \( \rho \) will be the zero vector) and the asymptotic distribution will simply be a (central) chi-squared.
References


Wald, A., 1943. Tests of statistical hypotheses concerning several parameters when the number of observations is large. Trans. Am. Math. Soc. 54, 426–482.