On the Advantage of Haplotype Analysis in the Presence of Multiple Disease Susceptibility Alleles

Richard W. Morris* **and Norman L. Kaplan**

Biostatistics Branch, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina

We investigated the effect of multiple susceptibility alleles at a single disease locus on the statistical power of a likelihood ratio test to detect association between alleles at a marker locus and a disease phenotype in a case-control design. Using simplifying assumptions to obtain the joint frequency distribution of marker and disease locus alleles, we present numerical results that illustrate the impact of historical variation of initial associations between marker alleles and susceptibility alleles on the power of a likelihood ratio test for association. Our results show that an increase in the number of susceptibility alleles produces a decrease in power of the likelihood ratio test. The decrease in power in the presence of multiple susceptibility alleles, however, is less for markers with multiple alleles than for markers with two alleles. We investigate the implications of this observation for tests of association based on haplotypes made up of tightly linked single-nucleotide polymorphisms (SNPs). Our results suggest that an analysis based on haplotypes can be advantageous over an analysis based on individual SNPs in the presence of multiple susceptibility alleles, particularly when linkage disequilibria between SNPs is weak. The results provide motivation for further development of statistical methods based on haplotypes for assessing the potential for association methods to identify and locate complex disease genes. Genet. Epidemiol. 23:221–233, 2002. © 2002 Wiley-Liss, Inc.

Key words: statistical power; likelihood ratio test; case-control design; single-nucleotide polymorphisms; linkage disequilibrium; complex genetic disease

*Correspondence to: Richard Morris, Biostatistics Branch, Mail Drop A3-03, National Institute of Environmental Health Sciences, PO Box 12233, Research Triangle Park, NC 27709-2233. E-mail: morrisr@niehs.nih.gov

Received for publication 11 March 2002; Revision accepted 6 May 2002
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/gepi.10200

INTRODUCTION

Studies of statistical methods for detecting and localizing genes contributing to risk for a complex disease depend on a number of simplifying assumptions. One assumption frequently employed in studies of methods that test for association between marker genotypes and disease phenotype is that a single susceptibility allele is present in the population at a disease locus [Clayton, 2000; Lazzeroni, 2001]. This assumption, embodied in the concept of a shared ancestral haplotype, is appropriate for populations in which chromosomes bearing a susceptibility allele trace their history to a common ancestor carrying the susceptibility mutation. Such a mutation may have been present in a small group of founders, or it may have arisen spontaneously. Alternatively, allelic variation at a complex disease locus may be extensive, with multiple susceptibility alleles of independent origin present in a population [Pritchard, 2001; Terwilliger and Weiss, 1998]. Because the genetic basis of complex human diseases is unknown, the hypothesis of a single susceptibility allele may be inappropriate when studying the prospects for association methods to identify complex disease genes.

An immediate concern raised by the hypothesis of multiple susceptibility alleles is the potential for limiting statistical power to detect association [Slager et al., 2000; Longmate, 2001]. It is not difficult to see how the presence of multiple, independent susceptibility alleles at a disease locus might adversely affect detection of association between marker genotypes and a disease phenotype. Suppose that a biallelic marker locus is tightly linked to a disease locus at which multiple susceptibility alleles arise by independent mutation. Susceptibility mutations occur on chromosomes bearing either marker allele in proportion to the frequencies at which the marker alleles occur in the population. If penetrances of genotypes that carry any of the susceptibility alleles are similar, then the frequency of a particular marker allele among affected individuals effectively depends on how often a susceptibility mutation occurs on a background bearing that marker allele. If the number of independent susceptibility mutations is large, then the frequencies of marker alleles among affected individuals will be very similar to the frequencies of marker alleles among unaffected individuals, making it difficult to detect an association between marker genotypes and a disease phenotype [MacLean et al., 2000; Schork et al., 2001]. Recombination, which reduces initial linkage disequilibrium between marker alleles and susceptibility alleles, would make the situation worse.

Here we investigate the consequences of multiple, independent susceptibility alleles on the power of a likelihood ratio test (LRT) to detect association in a case-control design. We studied power under alternatives in which linkage disequilibria between marker and susceptibility alleles arose from independent sampling of marker alleles by susceptibility mutations. Our primary conclusion is that in the presence of multiple susceptibility alleles an analysis based on haplotypes can be advantageous over an analysis based on individual single-nucleotide polymorphisms (SNPs), especially when linkage disequilibria between SNPs forming a haplotype are weak. This result provides motivation for further development of statistical methods based on haplotypes for assessing the potential for association methods to identify and locate complex disease genes.

METHODS

Likelihood Ratio Test

We studied the effect of multiple susceptibility alleles on the statistical power of an LRT to detect a difference in marker allele frequencies between cases and controls. We assumed a single population in which mating occurs at random with respect to marker-disease haplotypes, and computed power of the LRT under specific alternatives for a sample of 600 cases and 600 controls.

The LRT statistic for a test of the null hypothesis that marker allele frequencies are equal in case and control populations is given by $2[l(\mathbf{y};\,\hat{\boldsymbol{p}})) - l(\mathbf{y};\,\hat{\boldsymbol{p}}_0)]$, where $\mathbf{y}$ represents marker allele, haplotype, or genotype counts. $l(\mathbf{y};\,\hat{\boldsymbol{p}}0)$ is the value of the log-likelihood with separate maximum likelihood estimates of marker probabilities, $\hat{\boldsymbol{p}}$, in cases and controls, and $l(\mathbf{y};\,\hat{\boldsymbol{p}}_0)$ is the value of the log-likelihood under the null hypothesis with marker probability estimates, $\hat{\boldsymbol{p}}_0$, constrained to be equal in cases and controls. Under the null hypothesis, the LRT statistic is asymptotically distributed as central χ^2 with degrees of freedom equal to the number of constraints used to compute $l(\mathbf{y};\,\hat{\boldsymbol{p}}_0)$. Under an alternative hypothesis, the LRT statistic is asymptotically distributed as noncentral χ^2 with degrees of freedom as under the null.

The LRT statistic is based on a multinomial likelihood. For markers with two or more alleles and for haplotypes when haplotype phase was assumed known, the LRT statistic was computed using multinomial maximum likelihood estimates of marker allele or haplotype frequencies [Agresti, 1990]. When haplotype phase was assumed unknown, the LRT statistic was computed using an expectation-maximization algorithm to maximize a multinomial likelihood based on genotype frequencies under the assumption of Hardy-Weinberg equilibrium [Excoffier and Slatkin, 1995; Fallin et al., 2001]. We used a Newton-Raphson algorithm to maximize the complete-data likelihood and substituted the resulting maximum likelihood estimates into the observed-data likelihood, which has coupling and repulsion phase double heterozygotes confounded, to obtain the LRT statistic.

To compute the power of the LRT, we obtained the noncentral parameter of the χ^2 distribution for a specified alternative hypothesis by computing the LRT statistic using expected log-likelihoods in the expression for the LRT statistic. The resulting value of the LRT statistic was used as the noncentrality parameter of the noncentral χ^2. Expected log-likelihoods were formed by replacing counts $\mathbf{y}$ with their expected values, obtained as the product of sample size and the multinomial probabilities associated with a specific alternative hypothesis, as described below [for genetic applications, see Schaid, 1999; Longmate, 2001]. Using this noncentral parameter, the power for LRT under a specified alternative was obtained by computing the probability that a noncentral χ^2 distributed random variate exceeds the $(1-\alpha)$ percentile from a central χ^2 distribution with the same degrees of freedom. We used $\alpha = 0.01$ for all tests. The accuracy of this method of power computation was verified by computer simulations (results not shown).

To compare the power of a multiple-degree-of-freedom LRT based on haplotypes with the power based on individual SNPs, we defined an individual SNP procedure as the maximum LRT statistic among the set of LRT statistics computed individually for each SNP comprising a haplotype. Control of type I error associated with performing separate tests was accomplished by Bonferroni correction of the critical value. Results of permutation tests for individual SNP procedures suggested that Bonferroni correction at level $\alpha = 0.01$ is not conservative for 2 and 3 marker haplotypes (results not shown).

Specification of Alternative Hypotheses

The LRT has null hypothesis $\Pr\{m_i|A\} = \Pr\{m_i\}$ for all $i = 1, 2, \ldots, r$, where m_i indexes the i^{th} allele (or haplotype) among r marker alleles, and A denotes an affected phenotype. As shown in the Appendix, under random mating, $\Pr\{m_i|A\} = \sum_{j=0}^{t} \Pr\{m_i|g_j\} \Pr\{g_j|A\}$, where g_j represents allele $j = 0, 1, \ldots, t$ at the disease locus. The case for a single susceptibility allele is given by Akey et al. [2001] and Chapman and Wijsman [1998].

The null hypothesis is satisfied when all penetrances are identical (i.e., $\Pr\{g_j|A\} = \Pr\{g_j\}$ for all j). In this article, we limit consideration to additive penetrances of the following form: $\Pr\{A|g_0g_0\} = \beta_0$, $\Pr\{A|g_0g_j\} = \beta_0 + \beta_1$, and $\Pr\{A|g_ig_j\} = \beta_0 + 2\beta_1$, for $i, j = 1, 2, \ldots t$. Power computations were done using values of $\beta_0 = 0.01$ and $\beta_1 = 0.02$, which produce genotype relative risks of 3 or 5 for 1 or 2 copies of a susceptibility allele, respectively.

The null hypothesis is also satisfied when linkage equilibrium occurs between all marker-disease allele pairs (i.e., $\Pr\{m_i|g_j\} = \Pr\{m_i\}$ for all i and j). Since the joint marker-disease allele frequency distribution determines all pairwise linkage disequilibria, it is sufficient to specify the joint frequency distribution of marker and disease locus alleles along with a set of penetrances to compute the power for an alternative hypothesis, and thereby study the power of the LRT.

Single Historical Scenario

To illustrate the construction of a joint marker-disease allele frequency distribution for one possible historical scenario, we consider a single susceptibility allele and assume a large population in which a biallelic marker is segregating for alleles m_1 and m_2. Suppose that a susceptibility mutation, g_1, at a tightly linked disease locus occurs by chance on a chromosome bearing 1 of the 2 marker alleles, say, the m_2 allele. If the effects of recombination are minimal, then to a reasonable approximation, $\Pr\{m_2|g_1\} = 1$, which implies $\Pr\{m_2 \cap g_1\} = \Pr\{g_1\}$. Moreover, since g_1 initially occurred on a chromosome bearing m_2, we also have $\Pr\{m_1|g_1\} = 0$, which implies $\Pr\{m_1 \cap g_1\} = 0$. Among chromosomes bearing the nonsusceptibility allele g_0, we assume that marker allele frequencies are stable and write $\Pr\{m_i \cap g_0\} = \Pr\{m_i\} \Pr\{g_0\}$ for $i = 1, 2$.

Generalizing to multiple susceptibility alleles, the joint marker-disease allele frequency distribution has three types of elements:

$$\Pr\{m_i \cap g_0\} = \Pr\{m_i\} \Pr\{g_0\}$$

$$\Pr\{m_i \cap g_j\} = \left\{ \begin{array}{ll} \Pr\{g_j\} & \textit{if } g_j \textit{ occurs on a chromosome with marker } m_i \\ 0 & \textit{otherwise} \end{array} \right\}$$

where $j = 1, \ldots t$ indexes susceptibility alleles. Table I gives an example of one possible historical scenario for four marker alleles and four susceptibility alleles. This

TABLE I. Hypothetical Joint Distribution of a Marker Locus With Four Alleles and Disease Locus With One Nonsusceptibility Allele, g_0, and Four Susceptibility Alleles

	Disease locus alleles				
Marker alleles	g_0	g_1	g_2	g_3	g_4
m_1	$\Pr\{m_1\}\Pr\{g_0\}$	$\Pr\{g_1\}$	0	0	0
m_2	$\Pr\{m_2\}\Pr\{g_0\}$	0	$\Pr\{g_2\}$	0	$\Pr\{g_4\}$
m_3	$\Pr\{m_3\}\Pr\{g_0\}$	0	0	0	0
m_4	$\Pr\{m_4\}\Pr\{g_0\}$	0	0	$\Pr\{g_3\}$	0

[a]All LRT power computations, except those based on ancestral selection graph, assume $\Pr\{g_0\} = 0.80$ and equally frequent susceptibility alleles ($\Pr\{g_j\} = 0.20/t$ for $j = 1, 2, \ldots, t$).

hypothetical joint distribution is a reasonable approximation when four preexisting marker alleles are segregating in a population into which four susceptibility mutations were introduced at a tightly linked locus and increased in frequency quickly enough that recombination can be ignored. Omitting recombination results in complete linkage disequilibrium between marker alleles and susceptibility alleles, which produces the greatest opportunity for detecting an association. We consider the consequences of relaxing this assumption in the Discussion.

The distribution given in Table I differs from the initial conditions employed by Chapman and Wijsman [1998, their Table 1] and Akey et al. [2001] for a single susceptibililty allele. They fixed marginal marker allele frequencies at the time of sampling, whereas we fixed ancestral marker allele frequencies and allowed alternative mutational histories to influence marginal marker allele frequencies at the time of sampling, depending on how the mutation process distributed susceptibility alleles among marker alleles. Current marker allele frequencies can therefore be viewed as random variables whose values depend on the outcome of a random mutation process that associates susceptibility mutations with marker alleles. From this perspective, ancestral marker allele frequencies, $\Pr\{m_i\}$, can be thought of as expected values of current marginal marker allele frequencies over random mutational histories. In what follows, when we refer to marker allele frequencies, we mean ancestral marker allele frequencies.

The joint marker-disease allele frequency distribution is determined by marker allele frequencies, susceptibility allele frequencies, and the outcome of the historical process that associates particular susceptibility alleles with particular marker alleles as a consequence of independent mutations. For stable ancestral marker allele frequencies, we consider the consequences of alternative historical scenarios on LRT power when susceptibility alleles are equally frequent. This represents a worst-case scenario with respect to LRT power, because disease locus genotypes bearing a particular number of susceptibility alleles will be equally represented among cases. For a fixed number of susceptibility alleles, t, we set $\Pr\{g_0\} = 0.80$ and let $\Pr\{g_j\} = 0.20/t$ for all j.

Multiple Historical Scenarios

With two marker alleles and $t > 1$ susceptibility alleles, the number of susceptibility mutations associated with one of the marker alleles, say, m_1, is binomial with parameters $\Pr\{m_1\}$ and t. Each realization of the binomial process represents a particular mutational history, which has associated with it a joint marker-disease allele frequency distribution from which LRT power can be computed. Consequently, power can be viewed as a random variable that depends on initial associations between marker alleles and susceptibility mutations that occur with binomial probabilities. We generalize the binomial distribution for a two-allele marker to a multinomial distribution for an r-allele marker. With t susceptibility alleles, the multinomial distribution has parameters $\mathbf{t} = (t_1, t_2, \ldots, t_r)'$ and $\mathbf{p} = (p_1, p_1, \ldots, p_r)'$, where $\mathbf{t}$ is the vector of counts of susceptibility mutations associated with each of the r marker alleles, such that $\sum_{i=1}^{r} t_i = t$ and $\mathbf{p}$ is the vector of marker allele frequencies. For a fixed number of susceptibility mutations, the number of historical scenarios considered for LRT power computations is the number of outcomes for the multinomial distribution [Feller, 1957].

Expected LRT Power

Expected LRT power over all unobservable histories, denoted Λ, is given by

$$E_{\Lambda}[\Pr\{reject\, H_0\}] = \sum_{h \in \Lambda} \Pr\{reject\, H_0 | h\} \Pr\{h\},$$

where h indexes a particular historical realization. $\Pr\{reject\, H_0|h\}$ is the power of LRT conditional on a history, and $\Pr\{h\}$ is the multinomial probability for that history. For a given number of equally frequent susceptibility alleles, there can be substantial variation in conditional LRT power for different marker-susceptibility allele configurations. Consequently, we summarized the distribution of conditional LRT power resulting from historical variation by computing the probability of obtaining an LRT with power of at least 0.80. The quantile statistic is

$$\Pr\{LRT\, power \geq 0.80\} = \sum_{h \in s(\Lambda)} \Pr\{h\},$$

where $s(\Lambda)$ is the subset of histories for which conditional LRT power is at least 0.80. We computed LRT power for every history and summed the multinomial probabilities of histories for which conditional LRT power was 0.80 or greater.

Nonuniform Susceptibility Allele Frequencies

We relaxed the assumption of equally frequent susceptibility alleles by using a model employed by Pritchard [2001] to study properties of susceptibility alleles that originate by mutation and are subject to drift and weak purifying selection. The model, known as the ancestral selection graph (ASG), was originally described by Krone and Neuhauser [1997] and Neuhauser and Krone [1997], and represents an extension of neutral coalescent methods to accommodate weak selection [Nordborg, 2001]. We used software developed by Paul Fearnhead (http://www.maths.lancs.ac.uk/~fearnhea/software/) to generate samples of disease locus alleles from an ASG, with mutation and selection parameters set to values provided by Pritchard [2001]. In particular, we used ASG parameter values $4N_e s = 20$ and $4N_e \mu = 5$, where N_e is the effective population size, s is the amount of selection against susceptibility alleles, and μ is the mutation rate to susceptibility alleles (Pritchard's μ_s). The mutation rate from susceptibility to nonsusceptibility alleles (Pritchard's μ_N) was $10^{-2}\mu$.

For each ASG realization, representing a single historical scenario, a prespecified allele was designated the nonsusceptibility allele and all other alleles were designated susceptibility alleles. We used only realizations where the frequency of the nonsusceptibility allele was in the interval [0.50, 0.95]. The joint marker-disease allele frequency distribution used to compute LRT power was obtained by random assignment of susceptibility alleles to marker alleles. Conditional LRT power was computed for each ASG realization, and $\Pr\{LRT\ power \geq 0.80\}$ was estimated by the proportion of realizations with conditional LRT power at least 0.80.

RESULTS

Uniform Distribution of Multiple Susceptibility Mutations: 2, 4, and 8 Marker Alleles

Since, for a fixed number of marker alleles, equally frequent alleles yield the most powerful test [Chapman and Wijsman, 1998], we computed $\Pr\{LRT\ power \geq 0.80\}$ under a uniform distribution of marker alleles. Figure 1 shows that for a marker with 2, 4, or 8 alleles, $\Pr\{LRT\ power \geq 0.80\}$ declines as the number of susceptibility alleles increases. An interesting feature of Figure 1 is that with equally frequent marker alleles, $\Pr\{LRT\ power \geq 0.80\}$ does not decline as quickly for

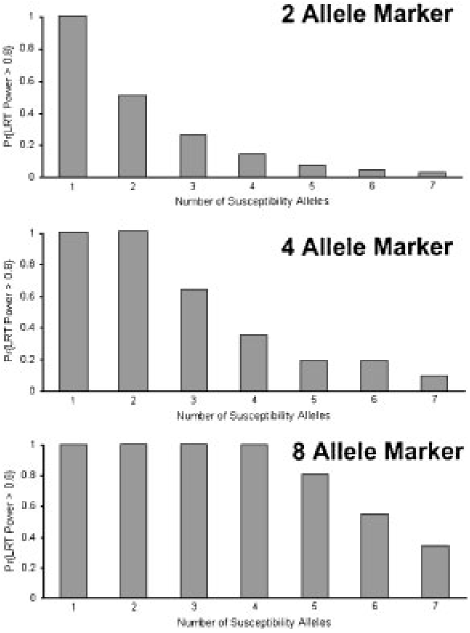

Fig. 1. $\Pr\{LRT\ power \geq 0.80\}$ for a marker with 2, 4, or 8 equally frequent alleles. Susceptibility alleles are equally frequent, and sample size is 600 cases and 600 controls.

a multiple allele marker as it does for a biallelic marker. Alternatively, for a given number of susceptibility alleles, Pr$\{LRT\ power \geq 0.80\}$ is greater with more marker alleles. This observation motivates power comparisons based on haplotypes.

Uniform Distribution of Multiple Susceptibility Mutations: Haplotypes

Figure 2 illustrates results for a 3-SNP haplotype for which each SNP has equally frequent alleles. Figure 2a represents uniformly distributed haplotype frequencies, and Figure 2b represents skewed haplotype frequencies, with 2 complementary haplotypes at high frequency (0.41) and 6 haplotypes at low frequency (0.03). Comparison of Pr$\{LRT\ power \geq 0.80\}$ among the three LRTs shows that phase-known and phase-unknown exhibit a power advantage over the individual SNP procedure. The power advantage is greater in Figure 2a, where linkage equilibrium holds, than in Figure 2b, which illustrates a skewed haplotype distribution. Similar results were obtained for a 2-SNP haplotype (results not shown).

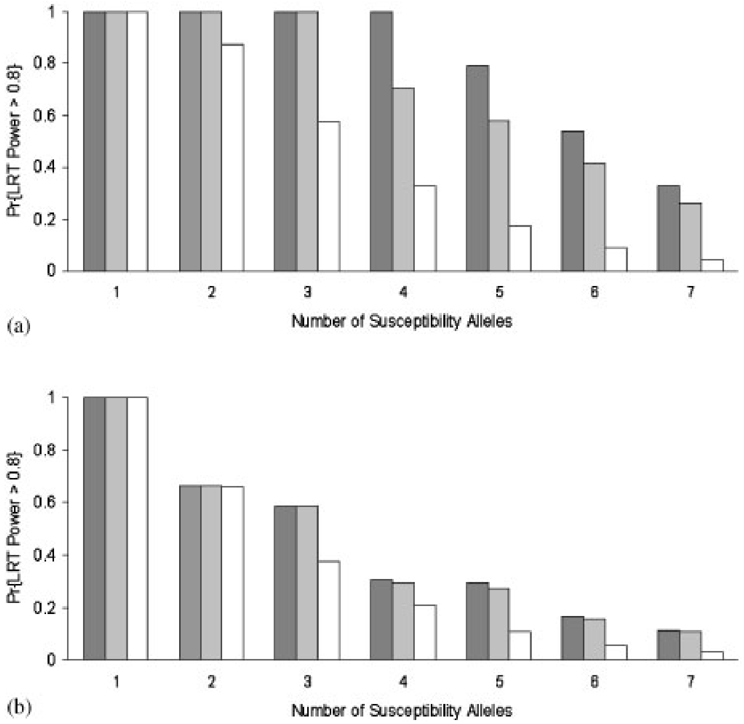

Fig. 2. Pr$\{LRT\ power \geq 0.80\}$ for a 3-SNP haplotype. Susceptibility alleles are equally frequent, and sample size is 600 cases and 600 controls. Two haplotype frequency distributions are illustrated. **a:** Equally frequent haplotypes. **b:** 0.41, 0.03, 0.03, 0.03, 0.03, 0.03, 0.03, and 0.41. Dark gray columns, phase known; light gray columns, phase unknown; open columns, individual SNP procedure.

To investigate extended haplotypes, we used haplotype data for six SNPs in the IL4RA gene for a sample of whites (170) and blacks (108) in the Chicago area [Table 1 in Ober et al., 2000]. The sample for whites has two haplotypes with frequencies 0.382 and 0.376 and 11 haplotypes with frequencies below 0.06. In contrast, the most frequent haplotype in blacks has a frequency of 0.194, and ordered haplotype frequencies exhibit a comparatively smooth decline. Consequently, haplotype frequencies in the sample for whites appear more skewed than in the sample for blacks.

In Figure 3, we use sample haplotype frequencies for six SNPs in the IL4RA gene as ancestral haplotype frequencies, and compare power for the phase-known LRT and the individual SNP procedure. For whites and blacks, Pr$\{LRT\ power \geq 0.80\}$ declines with an increase in the number of susceptibility alleles. For blacks, the decline in power is less, and the advantage of the haplotype procedure over the individual SNP procedure is greater than for whites. This result is consistent with observations on 3-SNP haplotypes in Figure 2 and observations on 2-SNP haplotypes (results not shown).

Nonuniform Distribution of Multiple Susceptibility Mutations

Figure 4 illustrates the impact of skewness of susceptibility allele frequencies on the value of Pr$\{LRT\ power \geq 0.80\}$ for 3-SNP haplotypes. For each of two 3-SNP haplotype distributions, two sets of LRT power estimates are plotted. One set, labeled "Ancestral Selection Graph", was computed directly from ASG output,

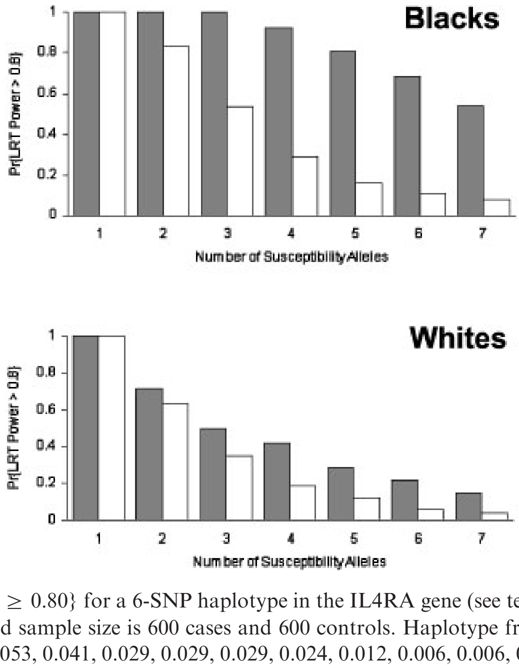

Fig. 3. Pr$\{LRT\ power \geq 0.80\}$ for a 6-SNP haplotype in the IL4RA gene (see text). Susceptibility alleles are equally frequent, and sample size is 600 cases and 600 controls. Haplotype frequency distribution for whites is 0.382, 0.376, 0.053, 0.041, 0.029, 0.029, 0.029, 0.024, 0.012, 0.006, 0.006, 0.006, and 0.006, and for blacks is 0.194, 0.148, 0.129, 0.111, 0.102, 0.074, 0.064, 0.056, 0.037, 0.028, 0.019, 0.019, and 0.018. Gray columns, phase known; open columns, individual SNP procedure.

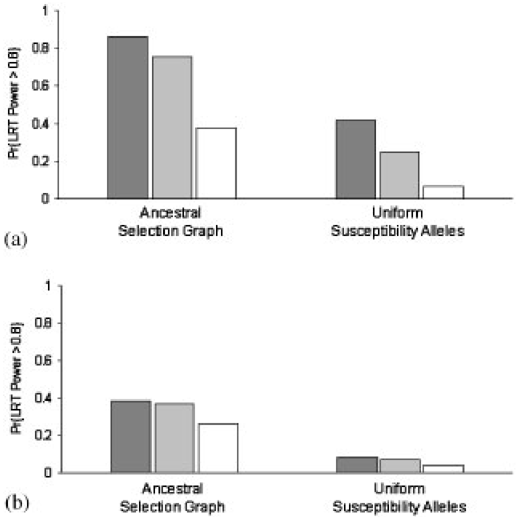

Fig. 4. Pr{*LRT power* ≥ 0.80} for a 3-SNP haplotype. For each history, the number of susceptibility alleles is given by ASG. Susceptibility allele frequencies are directly from ASG or are assumed to be equally frequent. Two haplotype frequency distributions are illustrated. **a:** Equally frequent haplotypes. **b:** 0.41, 0.03, 0.03, 0.03, 0.03, 0.03, 0.03, and 0.41. Dark gray columns, phase known; light gray columns, phase unknown; open columns, individual SNP procedure.

while the other set, labeled ‘‘Uniform Susceptibility Alleles’’, was computed using the observed number of alleles from each ASG realization but setting ‘‘Uniform Susceptibility Alleles’’ to be equally frequent when constructing the joint marker-disease allele frequency distribution. For each of the three LRTs, Pr{*LRT power* ≥ 0.80} is greater when computed directly from susceptibility allele frequencies generated under ASG than when computed using the number of alleles generated by ASG but forcing them to be equally frequent. This observation holds when all three SNPs are in pairwise linkage equilibrium and haplotypes are equally frequent (Fig. 4a) and when the 3-SNP haplotype marker distribution is skewed (Fig. 4b), although in the latter case power is reduced. We observed the same effect on power for ASG parameters $4N_es = 12$ and $4N_e\mu = 1$, and for 2-SNP haplotypes with both sets of parameters (results not shown).

DISCUSSION

Our study of the effects of multiple susceptibility alleles on the statistical power of an LRT to detect association is based on a variety of assumptions. Perhaps the most critical of these assumptions is the form of the joint frequency distribution of alleles at marker and disease loci. To highlight independent sampling of the background marker distribution, which is an essential feature of the effect of multiple susceptibility alleles, we assumed complete association between susceptibility alleles and marker alleles. This assumption produces a simple limiting form of the joint frequency distribution that retains initial marker-susceptibility allele associations resulting from independent mutations occurring on a background marker allele distribution. Scenarios where such a joint frequency distribution is reasonable require minimal recombination between marker and disease loci. This assumption is appropriate if marker and disease loci are tightly linked and susceptibility mutations are of recent origin. Young neutral susceptibility mutations may often be in low frequency; however, more common susceptibility mutations can also be young if allelic variation at the disease locus is due to a dynamic balance between mutation and drift with weak purifying selection [Pritchard, 2001]. We can relax the assumption of minimal recombination when constructing the joint marker-disease allele frequency distribution by allowing a fraction of initial association between marker allele and susceptibility allele to be distributed among other markers in proportion to ancestral marker allele frequencies. Computations performed using this surrogate for recombination dynamics resulted in reduced power of the LRT (results not shown). Relaxing the assumption of no recombination in this way, however, did not alter the qualitative results.

We have assumed throughout that different susceptibility mutations occur independently of one another in the history of the population. This assumption is key to our treatment of the problem of multiple susceptibility alleles. Most of our results, however, were obtained under two additional and admittedly artificial assumptions that can have an effect on the impact of multiple susceptibility alleles. One assumption is that susceptibility alleles are equally frequent, which represents a worst-case scenario. To investigate the effect of this special distribution, we used a plausible model for susceptibility alleles, the ASG, in which the resulting distribution is typically skewed with one or few frequent alleles and other less frequent alleles. In the extreme, a single susceptibility allele would be present. The effect of skewness is to mitigate the impact of multiple susceptibility alleles on power reduction of the LRT test. A second assumption is that all susceptibility alleles have equal penetrance. This assumption ensures that there is no differential representation of susceptibility alleles in cases due to differential penetrance. Relaxation of this assumption to a limiting case in which one susceptibility allele has high penetrance and all others have low penetrance could be expected to reduce allelic variation among cases and, like skewness in distribution, reduce the impact of multiple susceptibility alleles. It therefore follows that an ideal scenario for detecting association with a marker in the presence of multiple susceptibility alleles is for the disease locus to have a skewed distribution of susceptibility alleles, with the most frequent allele having the greatest penetrance. Whether this special scenario is representative of complex disease loci is unclear.

First and foremost, this study shows that the presence of multiple susceptibility alleles at a disease locus can reduce the power of the LRT test to detect association. This result was foreshadowed by Terwilliger and Weiss [1998] and illustrated for a biallelic marker in family-based association tests by Slager et al. [2000]. Our results extend these earlier results by showing that markers with multiple alleles also experience a decline in power with an increase in the number of susceptibility alleles. Consequently, if recent multiple susceptibility alleles are a feature of complex disease loci, then sample sizes larger than those indicated for a single susceptibility allele must be considered.

The results of this study suggest that in the presence of multiple disease susceptibility alleles, haplotype analysis can be advantageous over analyses based on individual SNPs. For a single susceptibility allele, Akey et al. [2001] reported that with equally frequent haplotypes made up of 2 or 4 biallelic sites, a chi-square test has greater power when based on haplotype than when based on individual SNPs. Their model involves recombination dynamics, but their conclusion is essentially the same as that of Chapman and Wijsman [1998], who showed that with equally frequent alleles, the power of the chi-square test is greater for a multiple allele marker than for a biallelic marker. The results of this investigation suggest that these conclusions can be extended to multiple susceptibility alleles at a disease locus.

In the presence of multiple susceptibility alleles, the power advantage of haplotype analysis over an individual SNP procedure depends on the degree of nonrandom association among component SNPs. With strong correlation among SNPs, marker variability is associated with a few common haplotypes. In this case, as our results suggest, the power advantage of haplotype analysis can be minimal or lost. Alternatively, if linkage disequilibrium between SNPs is weak, then statistical methods based on haplotypes may hold promise for identifying and locating disease genes if multiple susceptibility alleles are a general feature of complex disease genes.

ACKNOWLEDGMENTS

We thank Clarice Weinberg and Eden Martin for helpful comments on an earlier version of the manuscript.

APPENDIX

Denote an affected phenotype by A and let $\phi = \Pr\{A\}$. Let $m_i g_j$ index marker-disease haplotypes with alleles $i = 1, 2, \ldots, r$ at the marker locus, and alleles $j = 0, 1, \ldots, t$ at the disease locus. Let $m_i g_j / m_k g_l$ represent the two haplotypes of an individual. Then

$$\begin{aligned}\Pr\{m_i|A\} &= \sum_{k=1}^{r}\sum_{j=0}^{t}\sum_{l=0}^{t} \Pr\{m_i g_j / m_k g_l \cap A\}/\phi \\ &= \sum_{k=1}^{r}\sum_{j=0}^{t}\sum_{l=0}^{t} \Pr\{m_i g_j / m_k g_l\} \Pr\{A|m_i g_j / m_k g_l\}/\phi.\end{aligned}$$

Assume disease status depends only on genotype at the disease locus, and write $\Pr\{A|m_i g_j / m_k g_l\} = \Pr\{A|g_j g_l\}$. Further assume that genotypes are formed by random union of marker-disease haplotypes, so that $\Pr\{m_i g_j / m_k g_l\} = \Pr\{m_i g_j\} \Pr\{m_k g_l\}$. Then write gamete probabilities $\Pr\{m_i g_j\}$ as $\Pr\{m_i|g_j\} \Pr\{g_j\}$, and obtain

$$\Pr\{m_i|A\} = \sum_{k=1}^{r}\sum_{j=0}^{t}\sum_{l=0}^{t} \Pr\{m_i|g_j\} \Pr\{g_j\} \Pr\{m_k|g_l\} \Pr\{g_l\} \Pr\{A|g_j g_l\}/\phi.$$

Sum over marker alleles k and note that random union of marker-disease haplotypes implies $\Pr\{g_k g_l\} = \Pr\{g_k\}\Pr\{g_l\}$. Summation over disease alleles l gives the result:

$$\Pr\{m_i|A\} = \sum_{j=0}^{t}\sum_{l=0}^{t}\Pr\{m_i|g_j\}\Pr\{g_j g_l|A\} = \sum_{j=0}^{t}\Pr\{m_i|g_j\}\Pr\{g_j|A\}.$$

REFERENCES

Agresti A. 1990. Categorical data analysis. New York: John Wiley & Sons. p 48.

Akey J, Jin L, Xiong M. 2001. Haplotypes vs. single marker linkage disequilibrium tests: what do we gain? Eur J Hum Genet 9:291–300.

Chapman NH, Wijsman EM. 1998. Genome screens using linkage disequilibrium tests: optimal marker characteristics and feasibility. Am J Hum Genet 63:1872–85.

Clayton D. 2000. Linkage disequilibrium mapping of disease susceptibility genes in human populations. Int Stat Rev 68:23–43.

Excoffier L, Slatkin M. 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921–7.

Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D, Schork NJ. 2001. Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. Genome Res 11:143–51.

Feller W. 1957. An introduction to probability theory and its applications. 2nd ed. New York: John Wiley & Sons, Inc. p 36.

Krone SM, Neuhauser C. 1997. Ancestral processes with selection. Theor Popul Biol 51:210–37.

Lazzeroni LC. 2001. A chronology of fine-scale gene mapping by linkage disequilibrium. Stat Methods Med Res 10:57–76.

Longmate JA. 2001. Complexity and power in case-control association studies. Am J Hum Genet 68: 1229–37.

MacLean CJ, Martin RB, Sham PC, Wang H, Straub RE, Kendler KS. 2000. The trimmed-haplotype test for linkage disequilibrium. Am J Hum Genet 66:1062–75.

Neuhauser C, Krone SM. 1997. The genealogy of samples in models with selection. Genetics 145:519–34.

Nordborg M. 2001. Coalescent theory. In: Balding D, Bishop M, Cannings C, editors. Handbook of statistical genetics. Chichester: John Wiley & Sons, Ltd. p 179–212.

Ober C, Leavitt SA, Tsalenko A, Howard TD, Hoki DM, Daniel R, Newman DL, Wu X, Parry R, Lester LA, Solway J, Blumenthal M, King RA, Xu J, Meyers DA, Bleecker ER. 2000. Variation in the interleukin 4-receptor alpha gene confers susceptibility to asthma and atopy in ethnically diverse populations. Am J Hum Genet 66:517–26.

Pritchard JK. 2001. Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69:124–37.

Schaid DJ. 1999. Likelihoods and TDT for the case-parents design. Genet Epidemiol 16:250–60.

Schork NJ, Fallin D, Theil B, Xu X, Broeckel U, Jacob HJ, Cohen D. 2001. The future of genetic case-control studies. In: Rao DC, Province MA, editors. Genetic dissection of complex traits. San Diego: Academic Press, p 191–212.

Slager SL, Huang J, Vieland VJ. 2000. Effect of allelic heterogeneity on the power of the transmission disequilibrium test. Genet Epidemiol 18:143–56.

Terwilliger JD, Weiss KM. 1998. Linkage disequilibrium mapping of complex disease: fantasy or reality? Curr Opin Biotechnol 9:578–94.