Evaluating Associations of Haplotypes With Traits

Daniel J. Schaid*

Department of Health Sciences Research, Mayo Clinic College of Medicine, Rochester, Minnesota

Haplotypes have played a major role in the study of highly-penetrant single-gene disorders, and recent evidence that the human genome has hot-spots and cold-spots for recombination have suggested that haplotype-based methods may play a key role in the study of common complex traits. This report reviews the motivation of using haplotypes for the study of the genetic basis of human traits, ranging from biologic function, to statistical power advantages of haplotypes, to linkage disequilibrium fine-mapping. Recent developments of regression models for haplotype analyses are reviewed, offering a synthesis of current methods, as well as their limitations and areas that require further research. Regression models provide significant advantages, such as the ability to control for non-genetic covariates, the effects of the haplotypes can be modeled, step-wise selection can be used to screen for a subset of markers that explain most of the association, haplotype × environment interactions can be evaluated, and regression diagnostics are well developed. Despite these strengths, the current regression methods tend to lack the sophisticated population genetic perspectives offered by coalescent and other similar approaches. Future work that links regression methods with population genetic models may prove beneficial. *Genet. Epidemiol*. © 2004 Wiley-Liss, Inc.

Key words: case-control; cladistic; coalescent; generalized linear models; Hardy-Weinberg Equilibrium; linkage disequilibrium; variance component

Grant sponsor: U.S. Public Health Service, National Institutes of Health; Grant number: GM65450.

*Correspondence to: Daniel J. Schaid, Ph.D., Harwick 7, Mayo Clinic, 200 First Street, S.W., Rochester, MN 55905.
E-mail: schaid@mayo.edu

Received 2 September 2004; Accepted 3 September 2004
Published online 10 November 2004 in Wiley InterScience (www.interscience.wiley.com)
DOI 10.1002/gepi.20037

INTRODUCTION

Haplotypes, the combination of closely linked alleles on a chromosome, play key roles in the study of the genetic basis of disease. These roles vary from biologic function to providing information about ancient ancestral chromosome segments that harbor alleles that influence human traits. The main purpose of this report is to review the rapidly expanding developments of statistical methods for the association of haplotypes with different types of traits. Before these methods are reviewed, it is worthwhile to review the following factors that motivate the use of haplotypes: biologic function, statistical power advantages, and linkage disequilibrium (LD) mapping.

BIOLOGIC FUNCTION

Until recently, genetic markers on haplotypes were widely spaced, and not likely to possess biological function. In contrast, many of the single nucleotide polymorphisms (SNPs) in current use may define mutations within functional DNA variations, such as exons or promoters, and are densely spaced, so that haplotypes composed of these types of markers can have more of a biological role. As emphasized by Clark [2004 (this issue)], the functional properties of a protein are often determined by how it is folded, which in turn is determined by the linear sequence of the amino acids; this linear sequence is determined by DNA variation on a haplotype. There is strong evidence that several mutations within a single gene in *cis* position (i.e., on the same haplotype) can interact to create a "super-allele" that has a large effect on the observed phenotype. Some examples in humans are a gene that influences intestinal lactase activity [Hollox et al., 2001], a gene responsible for human lipoprotein lipase [Clark et al., 1998], the HPC2/ELAC2 gene that increases the risk for prostate cancer [Tavtigian et al., 2001], and a gene that influences actions of catecholamines, which influence bronchodilation, and hence asthma [Drysdale 2000]. Hence, there are strong biological reasons why haplotypes can be important to study.

STATISTICAL ADVANTAGES OF HAPLOTYPES

Haplotypes composed of SNPs that may or may not be functional can sometimes provide greater power than single-marker analyses for genetic disease associations, due to the ancestral structure captured in the distribution of haplotypes [Akey and Xiong, 2001]. The literature on the relative efficiency of analyzing haplotypes versus single markers is complicated by differing assumptions about the number of trait loci, the number of alleles at the trait loci, and the amount of linkage disequilibrium among alleles from the marker and trait loci. Most reports have compared the maximum of single-locus statistics, with Bonferroni correction for the multiple tests, to a global test of haplotype associations. The following conclusions seem justified. For a quantitative trait, when the set of measured SNPs includes causative SNPs, single-locus tests are more powerful than haplotype-based tests when the number of causative SNPs is less than the number of haplotypes [Bader, 2001]. Intuitively, when associations are concentrated on a small number of SNPs, the maximum of multiple single-locus tests is likely to be more powerful than a global test that considers all haplotypes. Furthermore, for some simulations based on coalescent theory, single-locus analyses were more powerful than haplotype-based methods even when the markers were not causative, but rather in LD with a dialleic locus that influences a quantitative trait [Long and Langley, 1999]. In contrast, for case-control studies, haplotype-based methods can be more powerful than single-locus analyses when the SNPs are in LD with a causative diallelic locus [Akey and Xiong, 2001]. Both single-locus [Slager et al. 2000] and haplotype-based methods lose power when there are multiple alleles at a causative locus, but haplotype-based methods lose less power. In this situation, the power advantage for haplotype-based methods is greatest when the marker alleles are not in strong LD with each other, yet in strong LD with the causative alleles [Morris and Kaplan, 2002]. This situation is likely to occur when the ages of the marker variants are much older than the ages of the alleles at the causative locus, so that the markers have weak LD by the time of origin of the disease susceptibility alleles. How often this occurs is unknown, but the above studies suggest that haplotype methods may be more powerful for younger, and hence more rare, causative variants, in contrast to older more common causative variants.

A limitation of the above comparisons is that they have focused on two extremes, either using the maximum of single-locus statistics or using a global test for all haplotypes. When there are many haplotypes, there are many degrees of freedom, which can weaken power to detect associations. In contrast, a multi-locus statistic that simultaneously tests for the main effects of all loci, yet without regard to haplotype phase, can have greater power than both the single-locus method and the haplotype method. A multi-locus test allows for association across multiple loci, yet has fewer degrees of freedom than the haplotype method. This feature has been observed for SNPs that are carefully chosen to ''tag'' common haplotypes [Chapman et al., 2003]. Further discussions of this multi-locus approach are given by Clayton [Clayton et al., 2004 (this issue)]. More comprehensive evaluations of the relative efficiencies of single-locus tests, multi-locus unphased genotype tests, and haplotype-based methods would help to clarify these issues.

LINKAGE DISEQUILIBRIUM MAPPING

Haplotypes have played a key role in the study of simple Mendelian diseases, mainly for fine-mapping, taking advantage of ancestral recombinations that trim haplotypes to informative small segments that harbor a disease locus. Because ancestral recombinations weaken associations, haplotype studies typically cover small chromosomal regions, ranging from tens of kilobases (kb's) to a few hundred thousand kb. Both haplotype association methods and LD fine-mapping methods take advantage of the fact that in the vicinity of a causative locus, haplotypes of diseased subjects tend to share more ancestry than haplotypes of unaffected subjects, and this excess sharing decreases with distance from the causative locus. Both approaches share the common goal of localizing the disease susceptibility locus. However, the means to accomplish this goal differs between the two approaches.

For LD fine-mapping, the main focus is either the distribution of complete marker haplotypes or pair-wise LD among the chromosomes from affected cases, with the chromosomes from controls playing a lesser role. The genealogy of the case chromosomes provides the most information on the position of the underlying trait locus. The controls are expected to share much less ancestry among themselves, and so the role of controls is often to provide information on the frequencies of haplotypes, or pair-wise marker LD, among haplotypes that are not ancestral to disease-bearing chromosomes. Some methods have focused on pair-wise associations of markers and use composite likelihoods [Terwilliger, 1995; Devlin et al., 1996; Xiong and Guo, 1997; Collins and Morton, 1998; Maniatis et al., 2004], because more ambitious complicated models are likely to be overly simplistic, and perhaps too stringent in their assumptions about population history [Devlin et al., 1996; Zhang et al., 2004]. Other methods account for the dependence among chromosomes from cases by modeling their ancestry [Hastbacka et al., 1992; Kaplan et al., 1995; Graham and Thompson 1998; Rannala and Slatkin 1998; Morris et al., 2000, 2002], others reconstruct the local genealogical tree [Lam et al., 2000], and others are based on haplotype sharing [McPeek and Strahs, 1999; Molitor et al., 2003a,b]. Some of the more recent methods are built on the modern population genetic coalescent theory. The parameters of these statistical models for the haplotype structure typically include recombination rates, mutation rates, the genealogy of the haplotypes, the location of the causative locus, the age of the causative allele(s), and demographic parameters, such as the size of the population and growth rates. These more complex models are attractive because they account for many of the population genetic features of interest. Because of the complexity of the likelihoods, markov chain monte carlo (MCMC) methods are used to fit the models. In practice, the more complex the model, the more parameters, and hence the longer the time to fit the models by MCMC. In theory, unphased haplotypes can be accounted for in the MCMC methods, treating them as latent variables. However, this increases the dimension of the parameters to estimate, slowing the convergence of MCMC. Because of this, most LD methods require haplotypes with known phase, which is typically accomplished by using the most likely pair of haplotypes per subject [for a review of statistical methods to infer haplotypes see Niu, 2004 (this issue)]. This practice is not satisfactory, because inferring the most likely pair of haplotypes, and then treating them as if they were directly observed, can result in a substantial loss of information and overly optimistic confidence intervals for parameter estimates [Morris et al., 2004]. The better approach is to directly analyze the unphased genotype data, accounting for haplotype ambiguity by statistical methods.

LD fine-mapping methods have achieved varying levels of success for Mendelian diseases of large genetic effect, such as Cystic fibrosis [McPeek and Strahs 1999; Morris et al., 2000, 2002; Liu et al., 2001; Molitor et al., 2003b], Friedreich's ataxia [Liu et al., 2001; Molitor et al., 2003b], Huntington's disease [Morris et al., 2000], Diastrophic dysplasia (Finland) [Hastbacka et al., 1992; Rannala and Slatkin 1998], Myoclonus epilepsy (Finland) [McPeek, 1999], hereditary hemochromatosis [Lam et al., 2000], and idiopathic distortion dystonia (Ashkenazi Jews) [Rannala and Slatkin 1998]. Recent attempts for the somewhat more complex Crohn's disease show promise [Conti and Witte, 2003]. The utility of the current arsenal of LD fine-mapping methods for common complex diseases, however, is not yet determined. Common diseases can be complicated by multiple genes, with both allelic and locus heterogeneity, both of which are likely when there are multiple rare variants of recent origin [Neale and Sham, 2004]. In addition, common disease can be complicated by phenocopies, as well as environmental risk factors that are stronger than the putative genetic risks, making it necessary to statistically account for environmental effects. Furthermore, complex diseases are studied with a variety of epidemiologic designs, such as case-control or cohort studies. Hence, although the population genetic principles of many LD fine-mapping methods are attractive, particularly because they provide an estimate of the location of the causative locus, the value of this current arsenal of statistical methods for complex traits is not yet known.

STATISTICAL METHODS FOR ASSOCIATION OF HAPLOTYPES WITH TRAITS

In contrast to the LD fine-mapping methods, the development of methods to evaluate the association of haplotypes with traits has followed the more traditional biostatistical path, essentially treating haplotypes as categorical covariates. This path offers the advantage of using a wide variety of established statistical regression methods, with necessary extensions to account for unphased haplotypes. This review discusses many of these regression approaches, offering insights to their properties, and raising considerations for further research. A potentially important area of future development is to bring a closer link between the population genetic principles underlying LD fine-mapping and the more traditional biostatistical regression framework.

TRADITIONAL HAPLOTYPE ASSOCIATION METHODS FOR CASE-CONTROL STUDIES

If haplotypes are directly observed, then it is simple to compare the frequencies of haplotypes between cases and controls, using many of the statistical methods that are used to compare allele frequencies. When phase is unknown, the underlying haplotypes can be treated as missing data within the expectation-maximization (EM) algorithm [Excoffier and Slatkin, 1995]. This allows estimation of haplotype frequencies for the cases and controls, and construction of a likelihood ratio statistic to test equality of haplotype frequencies between cases and controls, $LRT{=}2(\ln L_{cases}{+}\ln L_{controls}{-}\ln L_{pool})$, where the log likelihoods (In L) are maximized separately for the group of cases, for the group of controls, and for the pool of all subjects. Some limitations of this approach are that the chi-square approximation for the distribution of the LRT may not be adequate when there are many haplotypes (hence sparse data), it lacks adjustment for environmental covariates, it is restricted to categorical outcomes, and it is assumed that the pairs of haplotypes are in Hardy Weinberg Equilibrium (HWE) proportions. This latter assumption is somewhat strong, because the random pairing of haplotypes implies that the genotypes at each locus are expected to be in HWE [Schaid, 2004]. For the group of cases, the causative locus can be in HWE, but only if the allelic effects are multiplicative on the genotype relative risk [Clayton, 1999]. If this is not true, then the amount of departure from HWE will be determined by the underlying genetic mechanism, and the marker alleles that are in strong LD with the causative allele will be dragged away from HWE.

The success of this approach, as well as the more general regression methods discussed below, depends on getting the "right size" haplotypes. If the haplotypes are too long, composed of many distant loci that have recombined with the causative locus, then the haplotypes will be composed of many random alleles, with many haplotypes, diluting associations with disease. Hence, an important consideration is to scan the haplotype for the sub-haplotype that has the strongest association with disease, or perhaps for the single causative SNP within a haplotype that explains the haplotype association. This is very much like the "haplotype method" [Valdes et al., 1997], which matches cases and controls according to the genotype at the primary locus, and examines residual associations among secondary loci, to determine if the primary locus explains all of the haplotype-disease association. This can also be accomplished by step-wise regression, as discussed below.

REGRESSION MODELS FOR HAPLOTYPES

It is worthwhile to recall that when haplotypes have known phase, the generalized linear model (GLM) that describes how the haplotypes influence the mean of the trait, but not the scale, can be expressed as $E[Y]{=}f(X'\beta)$, where Y denotes the dependent trait, the haplotypes, treated as independent variables, are coded into the X matrix, β denotes the effects of haplotype pairs, and f is a function that generalizes the usual linear regression. In GLM terminology, $g(\,)=f^{-1}(\,)$ is the "link" function. Example link functions are the identity for quantitative traits and logit for binary traits. Although it is frequently assumed that the effects of haplotypes are additive, dominant and recessive relationships, or more general genetic models, can be incorporated into the X matrix. For example, Clayton has emphasized that haplotype effects statistically represent higher-order interactions among alleles on the same chromosome [Clayton and Jones, 1999], suggesting that stepwise methods can be used to evaluate the role of these types of interactions [Cordell and Clayton, 2002], and that ignoring interactions can sometime increase power to detect associations [Clayton et al., 2004].

An advantage of considering a GLM is that many regression methods are special cases, including logistic regression for binary traits, linear regression for quantitative traits, and parametric survival models for age of onset. For many complex traits, age of onset tends to occur earlier for those that have a genetic etiology. Hence, for cohort studies, it is critical to account for censored age of onset. Simple extensions also allow for the semiparametric proportional hazards model.

Furthermore, regression models provide a flexible means to adjust for environmental factors, and to evaluate haplotype x environment interactions. A key point is that when haplotype phase is known, the usual GLM considers the distribution of the trait *conditional* on the haplotypes and other covariates, and so in this case the distribution of the haplotypes does not impact the usual GLM.

Because the X matrix plays a critical role in the power to detect associations, we use a simple example of a binary trait and two loci, denoted A and B, each dialleic, to illustrate how interactions come into play, and their influences on haplotype effects. For this example, there are four two-locus haplotypes (denoted *AB, Ab, aB*, and *ab*). Assume that phase is known, so that the double heterozygotes are distinguishable, giving ten two-locus genotypes, as illustrated in Table I. For these ten genotype categories, there are nine degrees of freedom. We can simply parameterize the genotype frequencies in terms of nine odds ratios, using genotype *ab*/*ab* as the baseline. Or, we can consider nine parameters for a fully saturated logistic regression model, comprised of main effects and interactions. We shall consider two types of models: (1) a "locus" model that partitions allelic effects into within and between loci, and (2) a "haplotype" model that considers the main effects of haplotypes and their pair-wise interactions.

For the locus model, first consider a logistic regression model for the effect of the A locus. Using the genotype *a*/*a* as the baseline, we can use x_A to count the number of A alleles, and use $u_{A/a}$ to indicate whether a subject is the heterozygote *A*/*a*. Then, the logit model for the marginal effect of the A locus can be written as

$$\text{logit} = \alpha_o + \alpha_A x_A + w_A u_{A/a},$$

where α_o is the baseline parameter (which we shall ignore), α_A is the main effect of locus A, and w_A measures the departure from additive effects of the A allele, the interaction within locus A (e.g., dominant/recessive effects). We can express a similar marginal model for the B locus, with appropriate subscripts to denote the main effect and the within-locus interaction of locus B. So far, we have four parameters. When considering both loci jointly, there are four additional interactions described by the products of the x and u covariates: the interaction of the main effects, $x_A x_B$, the interactions of the main effects and the within-locus interactions ($x_A u_{B/b}$ and $x_B u_{A/a}$), and the within-by-within interaction ($u_{A/a} u_{B/b}$). Finally, there is the phase effect that distinguishes the risk between the double heterozygotes. Let v indicate whether a subject has the genotype *AB*/*ab*. Then, the fully saturated logistic model can be expressed as

$$\begin{array}{ll} \text{logit} = & \\ \alpha_o + \alpha_A x_A + \alpha_B x_B & \text{main effects} \\ \;+w_A u_{A/a} + w_B u_{B/b} & \text{within} - \text{locus interactions} \\ \;+\delta_{A*B} x_A x_B & \text{main} \times \text{main interaction} \\ \;+\delta_{A*w_B} x_A u_{B/b} + \delta_{B*w_A} x_B u_{A/a} & \text{main} \times \text{within interaction} \\ \;+\delta_{w_A w_B} u_{A/a} u_{B/b} & \text{within} \times \text{within interaction} \\ \;+\delta_{phase} v & \text{phase effect} \end{array}$$

where δ denotes a parameter for interaction between loci, with the subscript illustrating the type of interaction. The design matrix for this

TABLE I. Design matrices for regression models: locus model for allelic main effects, interactions, and phase effect, and haplotype additive model (elements with 0 are left blank)

	Locus model											
	Main effects				Interactions							
			Within-locus		Main × main	Main × within		Within × within	Phase	Haplotype additive model		
2-locus genotypes	x_A	x_B	$u_{A/a}$	$u_{B/b}$	$x_A x_B$	$x_A u_{B/b}$	$x_B u_{A/a}$	$u_{A/a} u_{B/b}$	v	z_{AB}	z_{Ab}	z_{aB}
AB/*AB*	2	2			4					2		
AB/*Ab*	2	1		1	2	2				1	1	
AB/*aB*	1	2	1		2		2			1		1
AB/*ab*	1	1	1	1	1	1	1	1	1	1		
Ab/*aB*	1	1	1	1	1	1	1	1			1	1
Ab/*Ab*	2										2	
Ab/*ab*	1		1								1	
aB/*aB*		2										2
aB/*ab*		1		1								1
ab/*ab*												

model is given in Table I. To express this in matrix notation, let X be a 10×9 design matrix, and let γ be the vector of all 9 parameters, so that logit $=X\gamma$.

Now consider the haplotype model. Instead of partitioning the effects between and within loci, assume that we model the haplotypes, using the haplotype *ab* as the baseline. In this case, there are three main effects, β_{AB}, β_{Ab}, and β_{aB}, and there are six pair-wise interactions, for the six genotypes that are heterozygous for their pairs of haplotypes. Hence, this model also has 9 parameters. Let Z be the 10×9 design matrix for the haplotype main effects and interactions, and let β be a vector for the corresponding parameters. Since both the locus and haplotype models are fully saturated, they must give the same fit to the data. So, we have $X\gamma = Z\beta$, and solving for β, we find that $\beta = (Z'Z)^{-1}Z'X\gamma$, which means that the haplotype model is simply a linear reparameterization of the locus model.

A more interesting question arises when we drop the interaction terms from the haplotype model, by removing the columns of Z that correspond to the pair-wise interactions (see Table I for this Z matrix). In this case, the additive effects of haplotypes derived from $\beta = (Z'Z)^{-1}Z'X\gamma$ are

$$\begin{aligned}
\beta_{AB} &= \alpha_A + \alpha_B + \left[w_A \tfrac{2}{9} + w_B \tfrac{2}{9}\right] + \delta_{A*B} \tfrac{49}{27} + \delta_{A*w_B} \tfrac{10}{27} + \delta_{B*w_A} \tfrac{10}{27} + \left[\delta_{w_A w_B} \tfrac{1}{9}\right] + \delta_{phase} \tfrac{4}{27} \\
\beta_{Ab} &= \alpha_A + \left[w_A \tfrac{2}{9} + w_B \tfrac{2}{9}\right] + \delta_{A*B} \tfrac{4}{27} + \delta_{A*w_B} \tfrac{10}{27} + \delta_{B*w_A} \tfrac{1}{27} + \left[\delta_{w_A w_B} \tfrac{1}{9}\right] - \delta_{phase} \tfrac{1}{54} \\
\beta_{aB} &= \alpha_B + \left[w_A \tfrac{2}{9} + w_B \tfrac{2}{9}\right] + \delta_{A*B} \tfrac{4}{27} + \delta_{A*w_B} \tfrac{1}{27} + \delta_{B*w_A} \tfrac{10}{27} + \left[\delta_{w_A w_B} \tfrac{1}{9}\right] - \delta_{phase} \tfrac{1}{54}
\end{aligned}$$

The main point to notice is the bracketed terms that contribute constant coefficients to the haplotype effects (β-terms), implying that within-locus interactions do not distinguish the β-terms haplotype effects, as expected for the assumed additive effects of haplotypes. However, the locus main effects, the interaction of locus main effects, the interaction of locus main effects and within-locus interactions, and the phase effect, do not have constant coefficients across the three different β's, implying that all of these effects influence the haplotype main effects. Hence, by the additive haplotype model, we have used only three parameters to capture the influence of six parameters from the locus model, which could improve power by the reduced degrees of freedom. Conti and Gauderman [2004] propose Bayes model averaging to determine if the interaction forms, and phase effect, should be kept in the model.

REGRESSION MODELS FOR UNPHASED HAPLOTYPES

To account for unphased haplotypes, some investigators have used statistical methods to infer the most likely haplotype pair per subject, and then use these inferred haplotypes as if they were observed. This approach does not account for the discarded haplotype pairs that are possible, and if LD is not strong, there can be substantial loss of information [Schaid, 2002]. In fact, using only the most likely haplotypes introduces measurement error into the X matrix, resulting in biased estimates of haplotype effects [Zhao et al., 2003], and possibly increased error in the estimated parameters [Tanck et al., 2003]. A more appealing approach is to use all possible pairs of haplotypes that are consistent with the observed marker data.

For unphased haplotypes, we can still use the general form of a GLM, but now we need to account for haplotype ambiguity by modeling the probabilities of the possible haplotype pairs per subject. To illustrate these methods, we shall use G to denote the unphased multilocus genotypes for a subject and $H = \{h_1, h_2\}$ to denote a particular pair of phased haplotypes. Assuming a prospective likelihood, the contribution to the likelihood by the *ith* subject can then be expressed as

$$L = \sum_{H \in G} P\{Y | X_e, X_g(H), \beta\} P(H), \qquad (1)$$

where the sum is over all possible pairs of haplotypes consistent with the observed genotypes, X_e represents the environmental covariates, $X_g(H)$ is the genetic covariate, determined by the numeric coding of each pair of haplotypes, β is the vector of regression coefficients (including effects of X_e, $X_g(H)$, and possibly their interactions), and $P(H)$ is the prior probability of haplotype pair H. Using this general formulation, one can use a GLM to model the conditional probability of the trait given covariates, $P\{Y | X_e, X_g(H), \beta\}$. A summary of reports that fall under this general scheme is given in Table II, illustrating the type of trait addressed, and whether

TABLE II. Regression methods for haplotype effects on traits

Trait	Likelihood	HWE	Covariates	Reference
Binary	Prospective	Pool	No	[Chiano and Clayton, 1998]
Binary	Prospective	Pool	Yes	[Mander, 2001]
Binary/quantitative	Prospective	Pool	Yes	[Tregouet et al., 2002]
Quantitative	Prospective	Pool	Yes	[Mander, 2002]
Quantitative	Prospective	Pool	Yes	[Tanck et al., 2003]
Binary	Prospective	Controls	Yes	[Zhao et al., 2003]
GLM	Prospective	Pool	Yes	[Lake et al., 2003]
GLM	Prospective	Pool	Yes	[Seltman et al., 2003]
Binary	Prospective	Pool	Yes	[Stram et al., 2003]
Binary	Retrospective	Controls	No	[Epstein and Satten, 2003]
Censored	Prospective	Pool	Yes	[Lin, 2004]
Binary	Prospective	Pool	Yes	[Durrant et al., 2004]

environmental covariates were included in the developed methods.

To fit the regression model of expression (1), it is necessary to maximize the log-likelihood for all subjects over the regression parameters (β) and the parameters that describe the probabilities of haplotype pairs. Alternatively, estimating equations could be used [Zhao et al., 2003]. For censored data, the semiparametric proportional hazards model can be used, which additionally requires estimation of the cumulative baseline hazard [Lin, 2004]. Because there are typically many haplotypes, it is often assumed that HWE holds for the haplotypes. In this case, $P(H)$ can be modeled as the product of haplotype probabilities; let q_h denote the population frequency for the *hth* haplotype (these must also be estimated). For a binary trait, the HWE assumption can be imposed on either the unaffected control subjects or the pool of all subjects; see Table II. For a haplotype that has a strong effect on disease status, HWE is not expected to hold among the diseased subjects, unless the effects of the haplotypes are multiplicative on the genotype relative risk, hence the reason for assuming HWE among only the controls. The impact of departures from HWE is discussed later.

To maximize the log-likelihood, the unphased haplotypes can be treated as missing data within the EM framework, giving rise to the posterior probability of the *jth* pair of haplotypes for the *ith* subject

$$P\{H_{i,j}|G,Y,X_e,X_g(H_{i,j}),\beta\} = \frac{P\{Y|X_e,X_g(H_{i,j}),\beta\}P(H_{i,j})}{\sum_{H\in G}P\{Y|X_e,X_g(H),\beta\}P(H)}. \qquad (2)$$

The E-step uses the current parameter estimates $(\hat{\beta}, \hat{q})$ in expression (2) to update the posterior probabilities. If there are sampling weights for the subjects, denoted w_i, then multiplying these times the posterior probability of expression (2) gives a weight, $w_{i,j}$, for each pair of haplotypes. These weights for haplotype pairs can then be used in the M-step to update $\hat{\beta}$ (by weighted regression). To update $\hat{q}_h$, the expected number of haplotypes of type h is computed by

$$E[\#h] = \sum_i \sum_{H_{i,j}\in G_i} w_{i,j} Count(h|H_{i,j}),$$

where $Count(h|H)$ counts the number of haplotypes of type h in the pair H, with a value of 0, 1, or 2. Then, the usual multinomial frequency estimate is used, $\hat{q}_h = E[\#h]/(2N)$, where N is the number of subjects. These types of regression methods are distributed in a package of routines called HaploStats, which run in the S-PLUS or R statistical packages. It should be recognized that simultaneously estimating both haplotype frequencies and haplotype effects will be limited to a relatively small number of loci, depending on the sample size, because of the many parameters to estimate.

HYPOTHESIS TESTING WITH SCORE STATISTICS

An advantage of the GLM framework is that it provides a means to construct score statistics to test the null hypothesis of no haplotype effects. These scores statistics, adjusted for environmental covariates, measure the covariance of the residuals of a GLM model that fits only the environmental covariates with the *expected* haplotype coding in the X matrix. The weights for the expected haplotype codings are the posterior probabilities of the haplotype pairs, given the observed

TABLE III. Example X matrix for unambiguous haplotype pairs (subjects 1 and 2) and enumerated haplotype pairs for ambiguous subject 3

			Haplotype X matrix (counts of haplotypes) (haplotype no.)						
Subject	$\frac{\text{Haplotype}-1}{\text{Haplotype}-2}$	Haplotype pair, no.	1	2	3	4	5	6	Posterior probability[a]
1	$\frac{111}{110}$	1,2	1	1	0	0	0	0	1
2	$\frac{011}{011}$	3,3	0	0	2	0	0	0	1
3	$\frac{010}{001}$	4,5	0	0	0	1	1	0	$P(4{,}5 \mid G)$
	$\frac{011}{000}$	3,6	0	0	1	0	0	1	$P(3{,}6 \mid G)$
3	$E[X]$[b]		0	0	$P(3{,}6 \mid G)$	$P(4{,}5 \mid G)$	$P(4{,}5 \mid G)$	$P(3{,}6 \mid G)$	

[a] Posterior probabilities are the probabilities of a pair of haplotypes, given the unphased genotypes (G) per subject; posterior probabilities sum to 1 per subject.

[b] $E[X] = \sum_{X \in G} X_g(H)P(H|G)$.

genotypes; these posterior probabilities are by-products of the EM algorithm used to estimate haplotype frequencies when ignoring the trait [Excoffier and Slatkin, 1995]. Under the null hypothesis, it is reasonable to assume HWE of the haplotypes. An example of an X matrix, assuming additive effects of haplotypes, for both unambiguous haplotype pairs and enumerated haplotype pairs with their posterior probabilities, is given in Table III. Subject 3 has ambiguous haplotypes; after listing all possible pairs, an X matrix can be constructed, along with the corresponding posterior probabilities for the rows of the X matrix. To illustrate the score statistic, let $\tilde{y}_i$ denote a fitted value from a GLM with only environmental covariates, and let $E[]$ denote an expectation over the posterior probability of haplotype pairs under the null hypothesis, given the observed marker data. For example, see Table III for $E[X]$ of subject 3. The score statistics can then be shown to be

$$U = \sum_{i=1}^{N} \frac{(y_i - \tilde{y}_i)}{a(\phi)} E_i[X],$$

where $a(\phi)$ scales the distribution, with a value of σ^2_{mse} for the normal distribution, and a value of 1 for binomial and Poisson distributions [Schaid et al., 2002]. The variance of this score vector accounts for the ambiguous haplotypes by use of Louis's score statistics with incomplete data [Louis, 1982]. An advantage of the score statistics is that they are rapid to compute, making it feasible to compute P values by simulations (i.e., randomly ordering $(y_i - \tilde{y}_i)$, and recomputing the statistics many times), which is more robust for sparse haplotypes than relying on asymptotic normality. An alternative to the score statistics is to simply use the expected X scorings of haplotypes, $E[X]$, in standard regression packages [Zaykin et al., 2002]. Although this latter approach does not account for the increased variance of the resulting statistic, limited simulations (Schaid, unpublished data) suggest that ignoring the additional variation in the X matrix is not harmful when the haplotype effects are not large. An advantage of this latter approach is that the average X matrix can be used for a wide variety of study designs, such as matched case-control studies.

DEPARTURE FROM HWE

Departure of the haplotype pairs from HWE can result from genotyping errors, selection, population stratification, or statistical chance. Because haplotype ambiguity is caused by heterozygous loci, departure from HWE in the direction of excessive heterozygosity can increase the error in haplotype predictions, yet excessive homozygosity does not [Fallin and Schork, 2000; Single et al., 2002]. Excessive heterozygosity can also lead to inflated Type-I error rates when analyzing ambiguous haplotypes with the prospective regression model [Lake et al., 2003].

REGRESSION MODELS FOR CASE-CONTROL STUDIES

A limitation of the prospective likelihood is that it does not account for how the sample was ascertained. For case-control studies, it should be clear that estimated haplotype frequencies will be biased when haplotypes are associated with disease, because cases are over-sampled. Furthermore, the regression coefficients for the haplotype

X matrix can be biased. This bias does not occur when phase is known, because in this situation, it has been shown that prospective and retrospective logistic models both give consistent estimates of the log-odds-ratio [Prentice and Pyke, 1979]; only the intercept is affected. The proof of this requires a saturated model for the distribution of the X covariate. For unphased haplotypes, however, the saturated distribution of haplotype pairs cannot be estimated. Rather, the probability of an underlying pair of haplotypes is typically modeled according to the haplotype frequencies and the odds ratios. This can cause the estimated regression coefficients to be biased. The amount of bias depends on how accurately the haplotypes can be predicted from the genotypes. The stronger the LD among the markers, the better the prediction of haplotypes from genotypes, resulting in little bias [Stram et al., 2003]. The ability of genotypes to predict haplotypes can be quantified in terms of the squared correlation coefficient between the true and predicted haplotype counts. If this correlation is at least 0.8, there is little bias in the coefficients, but for very small values, the bias can be substantial.

One way to correct for biased coefficients is to use sampling weights for the subjects, with weights based on the population disease prevalence [Stram et al., 2003]. Another way is to use a retrospective likelihood, modeling $P(G|Y)$ by summing over all possible pairs of haplotypes consistent with genotype G [Epstein and Satten, 2003]. This approach assumes HWE holds only for the controls, yet still uses both cases and controls to estimate haplotype frequencies, and of course haplotype odds-ratios. Some limitations of the current retrospective likelihood methods are that environmental covariates cannot be included (nor haplotype-environment interaction covariates), and they are not robust to departures from HWE among the controls when haplotype effects are dominant or recessive [Satten and Epstein, 2004], in contrast to some prospective methods that are robust to departures from HWE [Schaid et al., 2002; Zhao et al., 2003]. However, it is possible to introduce an additional parameter, a "fixation index," that accounts for the average departure from HWE across the different types of haplotype pairs, which reduces bias [Satten and Epstein, 2004]. Some of these methods, based on the retrospective likelihood, are implemented in the software CHAPLIN (case-control haplotype inference software).

MANY HAPLOTYPES, RARE HAPLOTYPES

When there are many haplotypes, we are often faced with the dilemma of how to account for the rare ones. Frequency estimates for the rare haplotypes can have large variances, due to sampling variation, as well as unknown phase [Fallin and Schork, 2000]. Furthermore, their corresponding regression parameter estimates will have very large variances, often leading to model instability. One approach is to not include rare haplotypes in the X matrix, yet this implicitly groups them into the baseline category. Another strategy is to group all rare haplotypes into a single category. This facilitates model fitting, yet makes it nearly impossible to interpret the regression coefficient for this heterogeneous grouping. A more appealing approach is to "shrink" the effects of each of the rare haplotypes. This shrinkage can be toward a common mean, with the effects of the rare haplotypes shrunk somewhat to the same degree as those haplotypes with which they are most similar. Alternatively, the effects of rare haplotypes can be shrunk toward the effects of the haplotypes that are most similar to the rare ones. This shrinkage has been accomplished in several different ways, but before we discuss some of these, it is worthwhile to review biased estimation for general linear models for quantitative traits, where shrinkage is towards a common mean. This offers a general framework, with many of the proposed methods variants of this general approach.

For the general linear model, $Y{=}X\beta{+}\varepsilon$, with the error terms having a multivariate normal distribution with mean zero and covariance matrix $V(\varepsilon) = \sigma_{\varepsilon}^2 V$, where matrix V is known, the usual least squares estimator is

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}Y.$$

In the presence of sparse data, the variance of $\hat{\beta}$ can be large; we can reduce this variance, at the price of introducing bias, by assuming that β is a random variable, and imposing structure on the covariance matrix of β. That is, we assume that β has a multivariate normal distribution with mean zero and covariance matrix $V(\beta) = \sigma_{\beta}^2 S$, where S is a matrix of known structure. This S matrix plays a key role in the shrinkage, and as we shall see, has a natural interpretation when modeling haplotypes. There is a large body of literature on this two-stage hierarchical modeling approach

[Leyland and Goldstein, 2001], with extensions to generalized linear models [Lee and Nelder, 1996], falling under the umbrella of generalized linear mixed models [McCulloch and Searle, 2001]. The Bayes estimator for β, allowing S to be singular, is

$$\tilde{\beta} = SX'(V + X'SX)^{-1}Y. \qquad (3)$$

Although it may not be unusual for S to be a singular matrix when it is used to measure similarity of haplotypes (discussed below), an intuitive way to see the shrinkage factor is to assume that S is non-singular, in which case expression (3) reduces to

$$\tilde{\beta} = (X'V^{-1}X + S^{-1})^{-1}X'V^{-1}Y.$$

This expression illustrates that S^{-1} serves to shrink the usual least squares estimator by inflating the "denominator" matrix. Ridge regression is a special case, when $S^{-1} = \lambda I$, where λ is a "penalty" term used for shrinkage.

For the analysis of haplotypes, the matrix S imposes structure on the covariances of the haplotype effects, as measured by β. How best to construct S, allowing for covariances determined by shared genealogy, requires further research. To date, intuition has been the major guide. Since covariances of haplotype effects are likely determined by the similarity of the haplotypes, a number of similarity measures have been proposed. Some simple similarities are a matching measure (having a value of 1 if two haplotypes match alike in state at all loci), a length measure (physical or genetic length of the longest contiguous interval of matching alleles), or a count measure (the number of alleles alike-in-state over all loci) [Tzeng et al., 2003]. An advantage of these simple similarity measures is that they are easy to implement in standard software. A further advantage of the count similarity is that it is not necessary to determine haplotype phase when using the hierarchical model for testing hypotheses. Tzeng et al. first found this property of the count similarity measure when comparing the average haplotype similarity for pairs of cases versus that for controls [Tzeng et al., 2003]. They showed that the count similarity measure has the special feature that unknown phase does not matter; their proposed statistic can be computed directly from unphased genotype data to get the same value of the statistic as if phase were known.

To see why phase does not impact the hierarchical model that uses the count similarity, assume for now that phase is known, and that we create the usual X matrix for additive haplotype effects, with values of 0, 1, or 2 in each column to count each of the haplotypes. Assume the linear model $Y=X\beta+\varepsilon$, where $\varepsilon \sim N(0, \sigma^2_\varepsilon I)$, and assume that $\beta \sim N(0, \sigma^2_\beta S)$, where S is the count similarity matrix. This can be translated to a variance component model where $Var(Y) = \sigma^2_\beta XSX' + \sigma^2_\varepsilon I$. The matrix XSX' does not depend on haplotype phase, which means that we can construct a likelihood ratio statistic to test the null hypothesis of no association of any of the haplotypes with the trait, $H_o : \sigma^2_\beta = 0$, without having to consider haplotype phase. To provide an intuitive explanation of why the count similarity does not depend on phase, let $a_{h_i,l}$ denote the allele on haplotype h_i for subject i (h_i=1, 2 for a subject) at locus l. When the count similarity is applied to a pair of subjects, each with two haplotypes, there are four pairs of haplotypes that contribute to the total measure,

$$T_{i,j} = \sum_{h_i=1}^{2}\sum_{h_j=1}^{2}\sum_{l=1}^{L} I\left[a_{h_i,l} = a_{h_j,l}\right].$$

Without changing the value of $T_{i,j}$ we can move the summation over loci to the far left, which means that at each locus we are creating a total count of allele matches between two subjects, and then summing over loci; this does not depend on haplotype phase.

To show why XSX' does not depend on phase, assume that there are K distinguishable haplotypes, and that all L loci are diallelic, with each locus having alleles labeled a and b. Let A be a $K \times L$ design matrix, such that a row of A is a pattern of 1's and 0's according to the presence or absence of allele a on a haplotype. The complement of the A matrix is the matrix B, with patterns of 1's and 0's according to the presence of allele b. The matrix AA' counts the number of matches of the a allele across loci, and the matrix BB' gives a similar count for the b alleles. Then, the total count of matched alleles for pairs of haplotypes is $S=AA'+BB'$, so that the matrix XSX' can be written as $XSX'=XAA'X'+XBB'X'$. The matrix XA simply counts the number of a alleles at each locus (columns of the matrix) for each subject (rows of the matrix), and does not depend on phase; XB gives a similar count for the b alleles. Because these matrices do not depend on phase, neither does XSX'. Although phase is not required for computing the likelihood ratio statistic to test whether $\sigma^2_\beta = 0$, phased haplotypes are required to estimate the best linear unbiased predictors,

$\tilde{\beta} = \sigma_{\beta}^2 SX'[\sigma_{\beta}^2 XSX' + \sigma_{\varepsilon}^2 I]^{-1}(Y - \bar{y})$, because SX' depends on phase.

Instead of shrinking haplotype effects towards a common mean, they can be shrunk such that similar haplotypes are forced to have similar effects. This can be accomplished by adding a penalty term to the usual log likelihood [Klerkx et al., 2003; Tanck et al., 2003),

$$\text{penalty} = -0.5\lambda \sum_i \sum_j s_{i,j} \left(\beta_i - \beta_j\right)^2.$$

This term penalizes haplotype effects that differ, despite high similarity measured by $s_{i,j}$. A difficulty implementing this approach is estimation of the penalty coefficient λ. Estimation by cross-validation appears promising, but requires more thorough evaluation.

CLADISTIC AND CLUSTERING OF HAPLOTYPES

An alternative approach to account for many unique haplotypes, as well as rare haplotypes, is to cluster haplotypes such that those within a cluster have the same effect on the trait. With much fewer clusters than haplotypes, power should increase, and interpretation of results should simplify. Because the relationships among present day chromosomes depend on their ancestral histories, and the trait-locus alleles are embedded in this history, it is hoped that this historical information can be used to define clusters of haplotypes with similar ancestry, and, hence, hopefully the same underlying alleles at the trait locus. Although the strategies to include this historical information border between purely haplotype association methods and LD fine-mapping methods that attempt to estimate the trait locus position, many are based conceptually on the coalescent process. The coalescent describes the tree structure that ties present-day haplotypes back to their most recent common ancestor, assuming that haplotype variation is caused by mutations, and not recombinations [Kingman, 1982]. Graphically, unique haplotypes, whether observed or ancestral, define the nodes of the tree, and a pair of haplotypes that have different alleles at only one locus are connected by a line, the connected set of haplotypes creates a tree. This lead to the development of cladistic analyses [Templeton et al., 1987, 1988, 1992, 2000; Templeton 1995, 1998], which uses a cladogram (an unrooted tree) to direct the search for trait-locus alleles. The cladistic analysis uses a sequential series of one degree-of-freedom tests to collapse the cladogram into a smaller one, effectively clustering haplotypes with similar effects. The potential advantage of the sequential testing approach is improved power over an omnibus test of all haplotypes. However, because collapsing is based on statistical testing, this depends on the size of the cladogram nodes (i.e., power to detect differences between nodes). Recently, this cladistic approach has been extended to the GLM framework, allowing for unphased haplotypes as described above [Seltman et al., 2003].

Analogous to the above cladistic analysis method, a logistic regression method for case-control data has been proposed, whereby standard hierarchical clustering is used to create a hierarchical tree of haplotypes [Durrant et al., 2004]. Likelihood ratio tests for logistic regression are then used to sequentially trim back the tree towards its root, creating fewer clusters, until further trimming gives a poor fit of the model. These methods focus on tests of association, rather than LD fine-mapping. To create the tree, a measure of "distance" between haplotypes is required. The authors used a measure that gives closer distance for haplotypes that share rare alleles than for those that share common alleles, since sharing rare alleles is expected to occur for haplotypes that share more recent common ancestry. Their measure of distance sums the contribution from different loci, instead of considering the joint frequency of alleles from multiple loci (i.e., haplotypes). Further improvements might be possible by using haplotype frequencies in a distance measure, which could potentially improve the information on the tree structure. Nonetheless, an advantage of the hierarchical clustering, followed by trimming, is a reduction in the number of haplotype groups, and a natural way to cluster the rare haplotypes. Although the proposed method required phased haplotypes [Durrant et al., 2004], unphased haplotypes could be easily analyzed by the following steps: (1) use the EM algorithm to enumerate the possible haplotypes, (2) cluster the distinguishable haplotypes, (3) create an X matrix for the clusters; an additive model would have a row for each possible pair of haplotypes, and the columns would serve to count the distinguishable clusters, just as the X matrix for haplotypes counts the haplotypes, (4) use the posterior probabilities to compute the average X per subject, (5) use the average X in regression models.

The success of the above cladistic and clustering methods depends to a large extent on the ability to construct an accurate cladogram, or hierarchical clustering. Most proposed methods first construct the cladogram ignoring the trait, and then use the constructed cladogram for analyses, as if it were constructed without error; current methods do not account for the statistical variation in the chosen cladogram, and how this impacts the association analyses. Furthermore, if recombination is allowed, the graph is no longer a simple tree, but rather a *network* of connected haplotypes [Griffiths and Marjoram, 1996; Nordborg and Tavare, 2002], which are difficult to model, even with MCMC methods, although some new approaches offer promise [Larribe et al., 2002]. Hence, some of the most challenging aspects of using cladograms to guide how to cluster haplotypes for association analyses are construction of cladograms, and accounting for their variation.

A potential limitation of the some of the above clustering methods, as well as some of the earlier discussed shrinkage methods for haplotype effects, is that they create the clustering, or the shrinkage *S* matrix, without regard to the association of the haplotypes with the trait. That is, they do not allow the clustering, or the *S* matrix, to adapt to the data, particularly to the position of the underlying trait locus. It would be much more informative to measure haplotype similarity *local* to the trait locus, than over the entire haplotype, particularly for long haplotype segments. Durrant et al. [2004] get around this problem by using a sliding window of SNPs, with windows determined by blocks of strong LD, and reconstructing the hierarchical tree within each window. The number of windows, and their sizes, influences the final results, because windows that are too long will include haplotypes with recombinations, which dilutes associations, yet many windows increases the stringency to reach statistical significance due to the need to correct for multiple testing. Further work is needed on how to best choose the size of the sliding window.

As an alternative to sliding windows, spatial clustering methods have been proposed, whereby haplotypes are clustered into units with similar effects, allowing for the clusters to be estimated, along with their effect sizes, and allowing for simultaneous estimation of the trait locus [Molitor et al., 2003b; Thomas et al., 2003]. These fine-mapping methods build on earlier work that used a similarity *S* matrix and Bayesian spatial modeling via conditional autoregressive (CAR) models [Molitor et al., 2003a]. The similarity matrix was defined as the length of a chromosome segment with identical markers around the trait locus, which must be estimated. Hence, the similarity matrix changes as the trait locus position is updated in the estimation process. The CAR approach shrinks the effect of a haplotype, say haplotype *h*, towards an average, where this average is defined in terms of the similarity of haplotype *h* with all other haplotypes; those that are most similar get the greatest weight. Some limitations of this approach are the need to estimate the effect of each haplotype (i.e., no clustering), and the implicit assumption of just one trait locus. If haplotypes fall into natural clusters, because different mutations at the trait locus occurred on different ancestral chromosomes, too much smoothing may occur. To overcome this, likelihood models have been developed that allow for multiple clusters, each with its own effect on the trait, hence allowing for multiple disease-causing variants. The number of clusters, and their effect sizes, are simultaneously estimated [Molitor et al., 2003b; Thomas et al., 2003].

An alternative Bayesian fine-mapping approach explicitly allows for multiple disease-causing mutations by clustering haplotypes while simultaneously estimating parameters of recombination rates, mutation rate, and location of the trait locus [Liu et al., 2001]. It is an advantage that the software for this method, BLADE, is easily available. A limitation, however, is that the number of clusters is fixed by the analyst. The authors assume that haplotypes that fall into the same cluster are mutually independent, conditional on the ancestral haplotype (i.e., a star genealogy for each cluster), which may not be robust if the number of clusters is misspecified.

Another Bayesian fine-mapping approach that attempts to build more of a coalescent model into the analysis, while allowing for multiple founding mutations and phenocopies, is the "shattered coalescent" [Morris et al., 2002]. By allowing branches of the haplotype genealogical tree to be removed during the estimation process, the tree can be "shattered" into subtrees, hence allowing for different ancestries among differing groups of haplotypes. Like many other methods, it initially required phased haplotypes, but extensions to allow for unphased genotypes (treating unknown phase as a latent variable) have increased its statistical efficiency, at the price of increased computation time [Morris et al., 2004]. Although it does not account for ascertainment (e.g., case-control studies), their simulations and application to Cystic fibrosis suggest that the estimation of the trait locus position is robust to ascertainment. The main assumption of this approach is that independent disease mutations must occur at the same locus. To allow for disease mutations at different loci, recombination would need to be included in the coalescent model, which is currently intractable for even small sample sizes (Andrew Morris, personal communication).

In summary, the spatial clustering approach uses the similarity matrix to assign haplotypes to clusters, whereas the Bayesian framework assigns haplotypes to clusters according to the estimated genetic parameters. Although it is not yet known how well these methods work for complex traits, nor how these general approaches compare to each other, they offer promising directions to link traditional regression models with haplotype clustering methods. Because of the complexity of the likelihood models for spatial clustering and the above Bayesian frameworks, MCMC methods are used to fit the models. Although MCMC methods are now common practice, they still require sophisticated statistical skills, with careful use of diagnostics to evaluate whether the models have converged. There remain substantial computational and statistical challenges, particularly for binary traits. Most current methods do not account for ascertainment, so it is not clear how well they will work for case-control studies of complex diseases. Most methods assume phased haplotypes. Unknown phase can be included as another latent random variable, but this increases the dimension of the parameters to estimate, which can slow the MCMC sampler as it moves throughout the target distribution (i.e., slow mixing), increasing the time to convergence. Another more subtle issue is the implication that a similarity matrix has on the MCMC properties. If random effects are correlated, then MCMC methods tend to have poor mixing. This is counter to what we are striving to achieve: using the similarity matrix to impose high correlations among similar haplotypes. Hence, a similarity matrix that imposes high correlations among haplotype effects can make it difficult to achieve model convergence within reasonable time limits. More research is required on how to handle unknown phase, the best types of similarity measures, and efficient implementations for MCMC methods.

MANY HAPLOTYPES, MANY DEGREES OF FREEDOM

When there are many haplotypes, the power to detect associations can weaken, due to the many degrees of freedom. One strategy is to compute a test statistic for each haplotype, and then use the maximum of these (i.e., smallest P value) to test for association, using the Bonferroni correction. While this approach may be most powerful when only one haplotype is strongly associated with the trait, its power is weakened when the association is spread out across multiple haplotypes. In this situation, a global test that considers all haplotypes simultaneously is more powerful. A global test can be derived from either a fixed effects model or a random effects model. As we illustrated above, the random effects model can also be formulated as a variance component model. Hence, an important issue is the relative power of these two approaches. If haplotypes are coded as additive effects in a fixed effects model, and the number of distinguishable haplotypes is denoted K, the degrees of freedom for the global F-statistic are K for the numerator and $N-K$ for the denominator, where N is the number of subjects. As K increases, the power can weaken because of the more stringent critical value. In contrast, no matter the value of K, the variance component model tests the null hypothesis $H_o: \sigma^2_\beta = 0$ versus the one-sided alternative hypothesis $H_o: \sigma^2_\beta > 0$, and so the likelihood ratio statistic has an asymptotic mixture distribution, analogous to using a one-sided test for a standard normal statistic. Because of this, it has been stated in several workshops and conferences that the variance component model is likely to be more powerful than the fixed effects model when there are many haplotypes.

To determine the relative power of fixed effects versus variance component models, I performed some simulations, albeit somewhat limited in scope. Using the coalescent approach, haplotypes composed of 10 SNPs were simulated by Hudson's MS program [Hudson, 2002], creating 13 haplotypes with frequencies illustrated in Figure 1. From this distribution, pairs of haplotypes were randomly sampled for 100 subjects, and the count sharing matrix, S, was constructed. For assumed values of σ^2_β ranging from 0 to 0.25, and fixed residual variance of $\sigma^2_\varepsilon = 1 - \sigma^2_\beta$, a vector of Y values was simulated according to a multivariate normal distribution with mean zero and covariance matrix $Var(Y) = \sigma^2_\beta XSX' + \sigma^2_\varepsilon I$. The heritability of the haplotypes is $h^2 = \sigma^2_\beta/(\sigma^2_\beta + \sigma^2_\varepsilon)$. From this data, both the score statistic for the

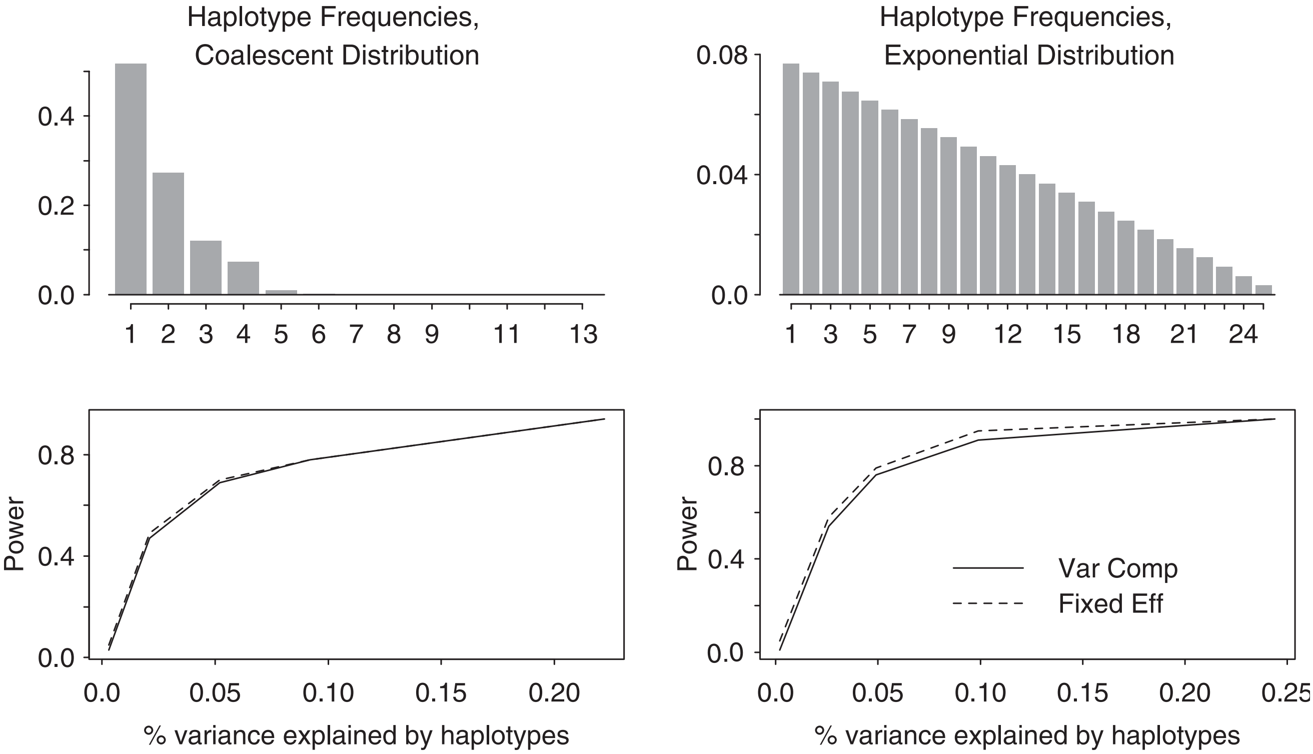

Fig. 1. Distribution of haplotypes by the coalescent simulations (top left) and an exponential distribution (top right), and their corresponding use for simulated power to compare fixed effects F-statistic versus variance component likelihood ratio statistic (bottom).

fixed effects model (allowing for unphased haplotypes) and the likelihood ratio statistic for the variance component model were computed, as well as their corresponding P values. This process was repeated 100 times to compute the power of these two methods. Results shown in Figure 1 illustrate that the power was almost identical for both methods. This is a bit surprising, given that the simulation process favored the variance component model, and the much fewer degrees of freedom for this test statistic. However, the fixed effects model may have performed well because there were not many degrees of freedom, a maximum of 13 haplotypes, yet with some random samples having fewer haplotypes due to the skewed distribution of haplotypes by the coalescent simulations (Fig. 1, top left). To further explore the impact of a larger number of haplotypes, I assumed an exponential distribution of 25 haplotypes, as illustrated in Figure 1 (top right). Somewhat surprisingly, the fixed effects model had slightly greater power than the variance component model, despite the many degrees of freedom (see Fig. 1, bottom right).

Since power is determined by the non-centrality parameter of the distribution of a test statistic, some insight to the relative power of fixed versus random effects for haplotype analyses can be gained by examining how the heritability of the haplotypes influences the non-centrality parameters. For the F-statistic, the non-centrality parameter is $\eta = NR^2/(1-R^2)$, where R^2 is the model multiple correlation coefficient. Since R^2 is interpreted as the percent of the variance of Y explained by the model, we can consider R^2 as the heritability of the haplotypes, h^2. To derive the non-centrality parameter for the variance component model, the general methods developed for pedigree variance component linkage analysis [Williams and Blangero, 1999] can be adapted to our situation. Here, the non-centrality parameter is $\eta = Nh^4 f(S)$, where $f(S)$ is a function that accounts for how the assumed similarity matrix influences the expected Fisher information. The main point is that the non-centrality parameter for the variance component likelihood ratio statistic depends on the *square* of the heritability, h^4, which can be dramatically smaller than h^2 that influences the non-centrality parameter for the F-statistic. Although this likely explains my simulation results, further research to evaluate how to best choose the haplotype similarity matrix is warranted, with more comprehensive comparisons between fixed effects and variance component models.

CONCLUSIONS AND FUTURE DIRECTIONS

Haplotypes will likely continue to play a major role in the study of the genetic basis of complex traits, particularly with our evolving understanding, and definitions, of "haplotype blocks" [Cardon and Abecasis, 2003]. The international HapMap project [Gibbs et al., 2003] will provide a wealth of data that will motivate the development and testing of new statistical methods, and application of the current arsenal of statistical methods to traits that are more complex than single major gene disorders should provide insights to their value. Experience with haplotype data from different populations should provide additional insights. Haplotypes composed of non-causative SNPs can be population-specific, depending on the population's ancestry and demography. This may be particularly relevant when a common disease is explained by common variants, where the age of the causative mutation allows for different recombination histories in different populations, compounded by different migration histories [Weiss and Clark, 2002; Neale and Sham, 2004]. In contrast, intragenic haplotypes, for which the causative variants are included, should provide greater power and reproducibility across different studies [Neale and Sham, 2004].

As outlined in this review, more work is needed to develop powerful methods to detect subtle associations of haplotypes with traits, yet robust to model misspecification, since most of our attempts to model the complexity of real data are likely to be overly simplistic. A quote from Albert Einstein captures this well: "As far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality" (Albert Einstein, 1879–1955). The regression approach offers several advantages: (1) it is possible to control for non-genetic covariates; (2) the effects of the haplotypes can be modeled; (3) step-wise selection can be used to evaluate each marker locus, or perhaps a set of loci making up a sub-haplotype, while controlling for the effects of all other marker loci; this allows one to screen for a subset of markers that explain most of the association [Valdes et al., 1997; Cordell and Clayton, 2002]; (4) haplotype × environment interactions can be evaluated [Lake et al., 2003]; (5) regression diagnostics are well developed. Linking GLM methods with population genetic models may prove beneficial.

ACKNOWLEDGMENTS

This review was greatly improved by the input from Andrew Morris, Glen Satten, Jason Sinnwell, Susan Slager, and Daniel Stram.

ELECTRONIC DATABASE INFORMATION

- CHAPLIN
  http://server2k.genetics.emory.edu/chaplin/download.html
- BLADE
  http://www.people.fas.harvard.edu/~junliu/index1.html
- SNPHAP
  http://cigmr.man.ac.uk/geneticanalysis/progs/snphap.html
- HaploStats
  http://www.mayo.edu/hsr/people/schaid.html
- MS
  http://home.uchicago.edu/~rhudson1/source/mksamples.html

REFERENCES

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

Bader J. 2001. The relative power of SNPs and haplotypes as genetic markers for association tests. Pharmacogenomics 2: 11–24.

Cardon L, Abecasis G. 2003. Using haplotype blocks to map human complex trait loci. Trends Genet 19:136–140.

Chapman JM, Cooper JD, Todd JA, Clayton DG. 2003. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. Hum Hered 56:18–31.

Chiano MN, Clayton DG. 1998. Fine genetic mapping using haplotype analysis and the missing data problem. Ann Hum Genet 62:55–60.

Clark AG. 2004. The role of haplotypes in candidate gene studies. Genetic Epidemiol 27:321–333 (this issue).

Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengard J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF. 1998. Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am J Hum Genet 63:595–612.

Clayton D. 1999. A generalization of the transmission/disequilibrium test for uncertain haplotype transmission. Am J Hum Genet 65:1170–1177.

Clayton D, Jones H. 1999. Transmission/disequilibrium tests for extended marker haplotypes. Am J Hum Genet 65:1161–1169.

Clayton D, Chapman J, Cooper J. 2004. The use of unphased multilocus genotype data in indirect association studies. Genet Epidemiol 27:415–428 (this issue).

Collins A, Morton NE. 1998. Mapping a disease locus by allelic association. Proc Natl Acad Sci USA 95:1741–1745.

Conti D, Gauderman J. 2004. SNPs, haplotypes and model selection in a candidate gene region: The SIMPle analysis for multilocus data. Genet Epidemiol 27:429–441 (this issue).

Conti D, Witte J. 2003. Hierarchical modeling of linkage disequilibrium: Genetic structure and spatial relations. Am J Hum Genet 72:351–363.

Cordell HJ, Clayton DG. 2002. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. Am J Hum Genet 70:124–141.

Devlin B, Risch N, Roeder K. 1996. Disequilibrium mapping: Composite likelihood for pairwise disequilibrium. Genomics 36:1–16.

Drysdale CM. 2000. Complex promoter and coding region b_2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. Proceed Nat Acad Sci 97:10483–10488.

Durrant C, Zondervan KT, Cardon LR, Hunt S, Deloukas P, Morris AP. 2004. Linkage disequilibrium mapping via cladistic analysis of single-nucleotide polymorphism haplotypes. Am J Hum Genet 75:35–43.

Epstein MP, Satten GA. 2003. Inference on haplotype effects in case-control studies using unphased genotype data. Am J Hum Genet 73:1316–1329.

Excoffier L, Slatkin M. 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921–927.

Fallin D, Schork N. 2000. Accuracy of haplotype frequency estimation for biallelic loci, via the Expectation-Maximization algorithm for unphased diploid genotype data. Am J Hum Genet 67:947–959.

Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, Ch'ang LY, et al.. 2003. The International HapMap Project. Nature 426:789–796.

Graham J, Thompson EA. 1998. Disequilibrium likelihoods for fine-scale mapping of a rare allele. Am J Hum Genet 63: 1517–1530.

Griffiths R, Marjoram P. 1996. Ancestral inference from samples of DNA sequences with recombination. J Comput Biol 3:479–502.

Hastbacka J, delaChapelle A, Kaitila I, Sistonen P, Weaver A, Lander E. 1992. Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. Nature Genet 2:204–211.

Hollox EJ, Poulter M, Zvarik M, Ferak V, Krause A, Jenkins T, Saha N, Kozlov AI, Swallow DM. 2001. Lactase haplotype diversity in the old world. Am J Hum Genet 68:160–172.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338.

Kaplan NL, Hill WG, Weir BS. 1995. Likelihood methods for locating disease genes in nonequilibrium populations. Am J Hum Genet 56:18–32.

Kingman JFC. 1982. The coalescent. Stochasatic processes and their Applications 13:235–248.

Klerkx AH, Tanck MW, Kastelein JJ, Molhuizen HO, Jukema JW, Zwinderman AH, Kuivenhoven JA. 2003. Haplotype analysis of the CETP gene: not TaqIB, but the closely linked -629C→ A polymorphism and a novel promoter variant are independently associated with CETP concentration. Hum Mol Genet 12:111–123.

Lake S, Lyon H, Silverman E, Weiss S, Laird N, Schaid D. 2003. Estimation and tests of haplotype-environment interaction when linkage phase is ambiguous. Human Heredity 55: 56–65.

Lam J, Roeder K, Devlin B. 2000. Haplotype fine mapping by evolutionary trees. Am J Hum Genet 66:659–673.

Larribe F, Lessard S, Schork NJ. 2002. Gene mapping via the ancestral recombination graph. Theor Popul Biol 62: 215–229.

Lee Y, Nelder JA. 1996. Hierarchical generalized linear models. J R Stat Soc B 58:619–678.

Leyland A, Goldstein H. 2001. Multilevel modelling of health statistics. New York: John Wiley & Sons, Inc.

Lin DY. 2004. Haplotype-based association analysis in cohort studies of unrelated individuals. Genet Epidemiol 26:255–264.

Liu JS, Sabatti C, Teng J, Keats BJ, Risch N. 2001. Bayesian analysis of haplotypes for linkage disequilibrium mapping. Genome Res 11:1716–1724.

Long AD, Langley CH. 1999. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. Genome Res 9:720–731.

Louis T. 1982. Finding the observed information matrix when using the EM algorithm. J R Stat Soc B 44:226–233.

Mander A. 2001. Haplotype analysis in population-based association studies. Stata J 1:58–75.

Mander A. 2002. Analysis of quantitative traits using regression and log-linear modeling when phase is unknown. Stata J 2: 65–70

Maniatis N, Collins A, Gibson J, Zhang W, Tapper W, Morton NE. 2004. Positional cloning by linkage disequilibrium. Am J Hum Genet 74:846–855.

McCulloch C, Searle S. 2001. Generalized, linear, and mixed models. New York: John Wiley & Sons, Inc.

McPeek MS. 1999. Optimal allele-sharing statistics for genetic mapping using affected relatives. Genet Epidemiol 16:225–249.

McPeek MS, Strahs A. 1999. Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping. Am J Hum Genet 65:858–875.

Molitor J, Marjoram P, Thomas D. 2003a. Application of Bayesian spatial statistical methods to analysis of haplotypes effects and gene mapping. Genet Epidemiol 25:95–105.

Molitor J, Marjoram P, Thomas D. 2003b. Fine-scale mapping of disease genes with multiple mutations via spatial clustering techniques. Am J Hum Genet 73:1368–1384.

Morris AP, Whittaker JC, Balding DJ. 2000. Bayesian fine-scale mapping of disease loci, by hidden Markov models. Am J Hum Genet 67:155–169.

Morris A, Whittaker J, Balding D. 2002. Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. Am J Hum Genet 70:686–707.

Morris RW, Kaplan NL. 2002. On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. Genet Epidemiol 23:221–233.

Morris AP, Whittaker JC, Balding DJ. 2004. Little loss of information due to unknown phase for fine scale linkage mapping with single-nucleotide-polymorphism genotype data. Am J Hum Genet 74:945–953.

Neale BM, Sham PC. 2004. The future of association studies: gene-based analysis and replication. Am J Hum Genet 75:353–362.

Niu T. 2004. Algorithms for inferring haplotypes. Genet Epidemiol 27:334–347 (this issue).

Nordborg M, Tavare S. 2002. Linkage disequilibrium: what history has to tell us. Trends Genet 18:83–90.

Prentice RL, Pyke R. 1979. Logistic disease incidence models and case-control studies. Biometrika 666:403–411.

Rannala B, Slatkin M. 1998. Likelihood analysis of disequilibrium mapping, and related problems. Am J Hum Genet 62:459–473.

Satten GA, Epstein MP. 2004. Comparison of prospective and retrospective methods for haplotype inference in case-control studies. Genet Epidemiol 27:192–201.

Schaid D. 2002. Relative efficiency of ambiguous vs. directly measured haplotype frequencies. Genet Epidemiol 23: 426–443.

Schaid D. 2004. Linkage disequilibrium testing when linkage phase is unknown. Genetics 166:505–512.

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. 2002. Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 70:425–434.

Seltman H, Roeder K, Devlin B. 2003. Evolutionary-based association analysis using haplotype data. Genet Epidemiol 25:48–58.

Single RM, Meyer D, Hollenbach JA, Nelson MP, Noble JA, Erlich HA, Thomson G. 2002. Haplotype frequency estimation in patient populations: the effect of departures from Hardy-Weinberg proportions and collapsing over a locus in the HLA region. Genet Epidemiol 22:186–195.

Slager SL, Huang J, Vieland VJ. 2000. Effect of allelic heterogeneity on the power of the transmission disequilibrium test. Genet Epidemiol 18:143–156.

Stram D, Pearce C, Bretsky P, Freedman M, Hirschhorn J, Altshuler D, Kolonel L, Henderson B, Thomas D. 2003. Modeling and E-M estimation of haplotype-specific relative risks from genotype data for case-control study of unrelated individuals. Hum Hered 55:179–190.

Tanck M, Klerkx A, Jukema J, DeKnijff P, Kastelein J, Zwinderman A. 2003. Estimation of multilocus haplotype effects using weighted penalized log-likelihood: analysis of five sequence variations at the cholesteryl ester transfer protein gene locus. Ann Hum Genet 67:175–184.

Tavtigian S, Simard J, Teng D, Abtin V, Baumgard M, Beck A, Camp N, Carillo AR, Chen Y, Dayananth P, Desrochers M, Dumont M, Farnham JM, Frank D, Feye C, Ghaffari S, Gupte JS, Hu R, Iliev D, Janecki T, Kort EN, Laity KE, Leavitt A, Leblanc G, McArthur-Morrison J, Pederson A, Penn B, Peterson KT, Reid JE, Richards S, Schroeder M, Smith R, Snyder SC, Swedlund B, Swensen J, Thomas A, Thranchant M, Woodland A-M, Labrie F, Skolnick MH, Neuhausen S, Rommens J, Cannon-Albright LA. 2001. A candidate prostate cancer susceptibility gene at chromosome 17p. Nature Genet 27:172–180.

Templeton A. 1998. Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history. Mol Ecol 7:381–397.

Templeton A, Crandall K, Sing C. 1992. A cladstic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. Genet 132:619–633.

Templeton A, Weiss K, Nickerson D, Boerwinkle E, Sing C. 2000. Cladistic structure within the human *Lipoprotein Lipase* gene and its implications for phenotypic association studies. Genetics 156:1259–1275.

Templeton AR. 1995. A cladstic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping or DNA sequencing. V. Analysis of case/control sampling designs: Alzheimer's disease and the apoprotein E locus. Genetics 140:403–409.

Templeton AR, Boerwinkle E, Sing CF. 1987. A cladistic analysis of phenotype associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in drosophila. Genetics 117:343–351.

Templeton AR, Sing CF, Kessling A, Humphries S. 1988. A cladistic analysis of phenotype associations with haplotypes inferred from restriction endonuclease mapping. II. The analysis of natural populations. Genetics 120:1145–1154.

Terwilliger JD. 1995. A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. Am J Hum Genet 56:777–787.

Thomas DC, Stram DO, Conti D, Molitor J, Marjoram P. 2003. Bayesian spatial modeling of haplotype associations. Human Hered 56:32–40.

Tregouet D, Barbaux S, Escolano S, Tahri N, Golmard JL, Tiret L, Cambien F. 2002. Specific haplotypes of the P-selectin gene are associated with myocardial infarction. Hum Mol Genet 11:2015–2023.

Tzeng JY, Devlin B, Wasserman L, Roeder K. 2003. On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. Am J Hum Genet 72: 891–902.

Valdes AM, McWeeney S, Thomson G. 1997. HLA class II DR-DQ amino acids and insulin-dependent diabetes mellitus: application of the haplotype method. Am J Hum Genet 60:717–728.

Weiss K, Clark A. 2002. Linkage disequilibrium and the mapping of complex human traits. Trends Genet 2002:19–24.

Williams J, Blangero J. 1999. Power of variance component linkage analysis to detect quantitative trait loci. Ann Hum Genet 63:545–563.

Xiong M, Guo SW. 1997. Fine-scale genetic mapping based on linkage disequilibrium: Theory and applications. Am J Hum Genet 60:1513–1531.

Zaykin D, Westfall P, Young S, Karnoub M, Wagner M, Ehm M. 2002. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered 53:79–91.

Zhang W, Collins A, Morton NE. 2004. Does haplotype diversity predict power for association mapping of disease susceptibility? Hum Genet 115:157–164.

Zhao L, Li S, Khalid N. 2003. A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. Am J Hum Genet 72:1231–1250.