Am. J. Hum. Genet. 70:124–141, 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

Heather J. Cordell and David G. Clayton

Department of Medical Genetics, University of Cambridge, Cambridge

A stepwise logistic-regression procedure is proposed for evaluation of the relative importance of variants at different sites within a small genetic region. By fitting statistical models with main effects, rather than modeling the full haplotype effects, we generate tests, with few degrees of freedom, that are likely to be powerful for detecting primary etiological determinants. The approach is applicable to either case/control or nuclear-family data, with case/control data modeled via *unconditional* and family data via *conditional* logistic regression. Four different conditioning strategies are proposed for evaluation of effects at multiple, closely linked loci when family data are used. The first strategy results in a likelihood that is equivalent to analysis of a matched case/control study with each affected offspring matched to three pseudocontrols, whereas the second strategy is equivalent to matching each affected offspring with between one and three pseudocontrols. Both of these strategies require parental phase (i.e., those haplotypes present in the parents) to be inferable. Families in which phase cannot be determined must be discarded, which can considerably reduce the effective size of a data set, particularly when large numbers of loci that are not very polymorphic are being considered. Therefore, a third strategy is proposed in which knowledge of parental phase is not required, which allows those families with ambiguous phase to be included in the analysis. The fourth and final strategy is to use conditioning method 2 when parental phase can be inferred and to use conditioning method 3 otherwise. The methods are illustrated using nuclear-family data to evaluate the contribution of loci in the *HLA* region to the development of type 1 diabetes.

Introduction

An important methodological issue in the identification of genes involved in complex disease is to distinguish between predisposing etiological variants and alleles at neighboring polymorphisms that may be in linkage disequilibrium (LD) with etiological variants but do not themselves have a direct involvement in causing disease. This task is considerably complicated by the high levels of LD observed between closely linked variants and by the fact that, even within a single gene, there may be more than one polymorphism involved in the disease. Once a genetic region involved in a complex disease has been localized (e.g., by use of linkage or association methods), a number of potentially causative sites may exist in the region, including a large number of single nucleotide polymorphisms (SNPs). A question of some interest is to determine which sites or combination of sites have a causal role in disease, and which show a disease association merely because of LD or because of their modifying effects on primary disease-causing polymorphisms in the region.

This question has received some attention in the study of HLA-associated diseases such as type 1 diabetes and rheumatoid arthritis, both of which show effects that map to the major histocompatibility complex (MHC). This region on chromosome 6 contains many genes involved in immunological response (Beck and Trowsdale 1999), but identification of the underlying disease genes is complicated by strong LD in the region. To distinguish between the primary (causal) and secondary effects in this region, several methods have been proposed. For example, the homozygous parent transmission/disequilibrium test (TDT) (Lie et al. 1999), motivated by the homozygous affected-sib-pair method (Robinson et al. 1993), considers transmissions of alleles from parents homozygous at the primary locus but heterozygous at a possible secondary locus, to affected offspring. A disadvantage of this test is that only parents who are homozygous at the primary locus contribute to the statis-

Received July 26, 2001; accepted for publication October 10, 2001; electronically published November 21, 2001.

Address for correspondence and reprints: Dr. Heather J. Cordell, University of Cambridge, Department of Medical Genetics, JDRF/WT Diabetes and Inflammation Laboratory, Cambridge Institute for Medical Research, Wellcome Trust/MRC Building, Addenbrookes Hospital, Hills Road, Cambridge, CB2 2XY, United Kingdom. E-mail: heather.cordell@cimr.cam.ac.uk

0002-9297/2002/7001-0013$15.00

tic, resulting in a considerable loss of information when the primary locus has more than two alleles. An alternative test is the haplotype method (Thomson et al. 1988; Valdes and Thomson 1997), which compares the relative frequencies of alleles at a secondary locus on haplotypes that are identical at a primary locus (or loci), in cases and controls. This method, which has been further developed by Li (2001), assumes that haplotypes of variants at *L* sites (loci) are available for a series of cases and controls. A similar method, in which odds ratios are calculated in case/control data for haplotypes with identical alleles at a primary locus (or loci), but differing alleles at a secondary locus, was used by Mignot et al. (2001). The haplotype method has also been applied to nuclear family data by incorporation of the method into a modified transmission/disequilibrium test (TDT) procedure (Cucca et al. 2001*a*) or by application of the method to affected family-based (AFBAC; Thomson 1995) controls (Zavattari et al. 2001). Construction of haplotypes in AFBAC controls is problematic, however, unless transmitted and untransmitted haplotypes can be unambiguously determined in every family, because discarding those families in which haplotypes are actually or potentially not inferable may induce a bias in the resulting estimated AFBAC haplotype frequencies.

A disadvantage of the haplotype method is that it may not be clear which haplotype background (i.e., which alleles at the primary locus or loci) should be used. If the test is repeated on a large number of backgrounds, a multiple-testing problem will result. This problem may be overcome by use of standard epidemiological procedures for analysis of case/control data, such as logistic regression, in which the additional contribution of the secondary locus is evaluated by comparison of a model in which the main effects and statistical interaction terms for both loci on a haplotype are modeled with one in which the main effects at the primary locus only are included. This is similar in spirit to the conditional extended transmission disequilibrium test (CETDT) proposed by Koeleman et al. (2000) for the analysis of family data.

A disadvantage of methods that focus on haplotypes as opposed to the multilocus (phased) genotypes of individuals is that these methods ignore any additional information contained in the combination of haplotypes present in an individual. Focusing on the haplotype as the unit of interest makes sense when one is aiming to localize a single etiological variant using LD methods under the assumption that the causative allele at the variant arose on a specific ancestral haplotype. In this case, the causative polymorphism itself may not even be present in the data set, although polymorphisms in varying degrees of LD with the causative polymorphism will be present. If, on the other hand, one has a collection of polymorphisms, several of which are likely to be etiological, it makes more sense to consider the overall combination of genotypes at the set of loci. In this way, we can allow for or indeed specifically test for dominance effects at any of the loci. For example, suppose we have two diallelic variants of interest in a region, variant 1 with alleles *a* and *A* and variant 2 with alleles *b* and *B*. Then denoting chromosomes (haplotypes) in the order paternal/maternal, there are 16 possible phased genotypes for an individual, which may be represented as *ab/ab*, *ab/aB*, *ab/Ab*, *ab/AB*, *aB/ab*, *aB/aB*, *aB/Ab*, *aB/AB*, *Ab/ab*, *Ab/aB*, *Ab/Ab*, *Ab/AB*, *AB/ab*, *AB/aB*, *AB/Ab*, and *AB/AB*. The full genotype model for the effects of these two loci would model the probability of being affected with disease as some function of 16 parameters corresponding to the 16 possible underlying genotypes. Assuming no parent-of-origin effects, we may assume that the disease risk for genotype *ij*/*kl* equals that for genotype *kl*/*ij*, and so we may model the probability of being affected with disease as some function of 10 parameters corresponding to the 10 phased genotypes (without distinguishing the parental origin of chromosomes) *ab/ab*, *ab/aB*, *ab/Ab*, *ab/AB*, *aB/aB*, *aB/Ab*, *aB/AB*, *Ab/Ab*, *Ab/AB*, and *AB/AB*. A model in which only variant 1 was important would model the disease probability as a function of just three parameters, corresponding to the genotypes *a/a*, *a/A*, or *A/A* at that locus. To test whether variant 2 is important once we have accounted for the effects of variant 1, we could therefore compare the fit of the 10-parameter model, which takes into account phased genotypes at both loci, to that of the 3-parameter model, which takes into account only genotypes at locus 1. To test whether the overall combination of variants 1 and 2 is important in disease we could compare the fit of the 10-parameter model to that of a single-parameter model in which the probability of being affected with disease does not depend on genotype at either locus.

The phased genotype model makes a distinction between the risks for genotypes *ab*/*AB* and *aB*/*Ab*. If we are willing to make the assumption that these two genotypes have the same disease risk—that is, that there are no "haplotype effects"—we may reduce the number of parameters to nine. This assumption is reasonable if the loci included in the model are the true disease-causing etiological variants. If the loci are, in fact, associated with disease because of LD with the true etiological variant(s), we would expect these "haplotype effects" to be non-negligible. Indeed, the existence of such effects would provide convincing evidence that a further locus is involved.

In the absence of haplotype effects, the data may be conveniently statistically represented in a linear model,

in which the probability of disease becomes a function of the genotypes at the two loci and of epistatic interactions between them:

$$\begin{aligned}\text{Prob(disease)} = f(&\beta_0 + \beta_1 I_{aA} + \beta_2 I_{AA} \\ &+\beta_3 I_{bB} + \beta_4 I_{BB} + \beta_5 I_{aA} I_{bB} \\ &+\beta_6 I_{aA} I_{BB} + \beta_7 I_{AA} I_{bB} + \beta_8 I_{AA} I_{BB}) \ ,\end{aligned}$$

where I_g is an indicator function taking the value 1 if an individual has genotype g and 0 otherwise, and $\beta_0, \beta_1, \ldots \beta_8$ are the nine parameters to be estimated. For this model, it is not necessary to know the phase of a doubly heterozygous individual. This is likely to be an advantage, since it avoids throwing away data for which the unphased genotypes are known but the haplotypes present (i.e., the phase information) cannot be determined. If, in fact, we are interested in narrowing down primary functional variants, rather than indirectly associated ones, we may reduce the model still further by ignoring epistatic interaction terms and fitting the model

$$\begin{aligned}\text{Prob(disease)} = f(&\beta_0 + \beta_1 I_{aA} \\ &+ \beta_2 I_{AA} + \beta_3 I_{bB} + \beta_4 I_{BB}) \ .\end{aligned}$$

Comparison of this model, in which effects at both variants are included, to the model $f(\beta_0 + \beta_1 I_{aA} + \beta_2 I_{AA})$, in which effects at locus 1 only are included, provides a test for the "main effects" of locus 2 while controlling for confounding at locus 1. This test has 2 df, corresponding to the difference in numbers of estimated parameters between the two models. This is considerably less than the 6-df test achieved when epistatic interactions are included or the 7-df test achieved when haplotype and epistatic effects are included. Similarly, the main effects test for the combined effect of both loci (against the null hypothesis that neither are involved) has 4 df, as opposed to 8 or 9 df when epistatic and/or haplotype effects are included. The degrees of freedom for the main effects test can be reduced still further by making additional assumptions, such as the absence of dominance, which would be equivalent to assuming that $\beta_4 = 2\beta_3$ and/or $\beta_2 = 2\beta_1$.

The reduced number of degrees of freedom possible in the "main effects" test suggests that for functional variants, this framework will provide the most powerful test of whether the loci are genuinely etiological. Given a small number of loci, these models can be examined by looking at the association between genotype and disease at each locus in turn, stratifying by genotype at the other loci. This would be similar in spirit to many of the previously proposed procedures, such as the haplotype method and homozygous-parent test. However, a more natural and convenient way to achieve the same result is to fit the models in a regression framework. Linear-regression models can be fit in most standard statistical packages. For case/control data the likelihood takes a form that leads to logistic-regression analysis (for unmatched cases and controls), whereas for family studies, the likelihood takes the form for conditional logistic regression (or matched case/control) analysis. Given a trio consisting of parents with a single affected offspring, and assuming haplotypes in the parents are known, we use the affected offspring as a case individual and construct three matched "pseudocontrol" individuals with phased genotypes constructed from the three other possible combinations of haplotypes that could have been transmitted from the parents (Self et al. 1991). This procedure was used by Thomas et al. (1995), who implemented the conditional logistic regression analysis via an empirical Bayes approach to deal with problems of multicolinearity and sparse data. Alternatively, Falk and Rubenstein (1987) proposed constructing, for each case, a single matched pseudocontrol whose phased genotypes consist of the haplotypes not transmitted from the parents to the affected offspring. However, the subsequent analysis proposed by these authors ignored the resulting matching in these data.

Previous implementations of conditional logistic regression–type methods for family data have assumed that haplotypes in the parents are either known or can be inferred from whatever offspring is produced (see the method 1 subsection, below). This may be true for highly polymorphic systems such as *HLA*, but it is unlikely to be true in general for a genetic region of interest. At the very least, we are likely to have to discard families for which this condition is not satisfied. Here, we propose three additional methods that allow us to make use of some or all such families. In the method 2 subsection, below, we show how to use families in which the case and one or more of the three potential pseudocontrols are able to resolve the parental phase. In the method 3 and method 4 subsections, below, we show how all typed families may be used, regardless of whether parental phase is resolvable, provided that we wish only to fit a model for the main effects of the loci of interest.

Methods

Case/Control Data

Suppose we have a sample of n_1 cases and n_2 controls, each genotyped at L polymorphisms (loci) within a gene

believed to be associated with disease. The likelihood of the observed data can be written as

$$\prod_{j=1}^{N} p_j^{I_j}(1-p_j)^{1-I_j} ,$$

where the product is over all $N = n_1 + n_2$ individuals, p_j represents the probability of individual j being a case rather than a control, and I_j is an indicator function taking the value 1 if individual j is a case and 0 otherwise. Following the standard statistical framework for generalized linear models (Nelder and Wedderburn 1972; McCullagh and Nelder 1989), we model p_j as

$$p = \frac{e^{\beta^T x}}{1 + e^{\beta^T x}}$$

or, equivalently, as

$$\text{logit}(p) = \ln\frac{p}{1-p} = \beta^T \mathbf{x} ,$$

where $\mathbf{x}$ is a vector that depends on the genotypes of the individual and β a vector of coefficients to be estimated. The length of the vector $\mathbf{x}$ and the coding scheme whereby $\mathbf{x}$ is related to the genotypes at the L polymorphisms determine the current model for the effects of these loci. Suppose we wish to model the effect of a single SNP locus. Then the possible genotypes at the SNP are 1/1, 1/2, and 2/2, which may be coded (for example) as $x_1 = -1, 0, 1$. The linear model is then

$$\text{logit}(p) = \beta_0 + \beta_1 x_1 ,$$

where β_0 (the intercept) and β_1 (the effect due to the SNP) may both be estimated. This model can be compared to the model

$$\text{logit}(p) = \beta_0 ,$$

where β_0 alone is estimated, using either a likelihood ratio or an efficient score test. In this way, we test whether the data is significantly better represented when the SNP is included in the model compared to when it is not in the model. This is equivalent to testing whether the coefficient β_1 is significantly different from 0. Note that the estimate β_0 is not biologically meaningful unless the sample is actually a population cohort, in which case β_0 estimates the log odds of being affected with disease in the population.

The $x_1 = -1, 0, 1$ coding scheme assumes that the effect of having two copies of allele 2 is twice that of having a single copy (i.e., there is no dominance effect). In this case, e^{β_1} represents the odds ratio for disease due to allele 2, and the test is equivalent to testing whether this odds ratio equals 1. Alternatively, the full effect of the SNP genotype can be modeled by specifying an additional variable x_2 (e.g., coded -0.5, 0.5, -0.5 for genotypes 1/1, 1/2, and 2/2) and comparing the full genotype model

$$\text{logit}(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

with

$$\text{logit}(p) = \beta_0 .$$

In this case, β_1 is a parameter representing the additive effect of allele 2 on the logit scale, and β_2 is a parameter representing the dominance effect of allele 2 over allele 1. If the polymorphisms in the gene are multiallelic rather than diallelic, the coding scheme can be extended to reflect this (e.g., see the coding schemes discussed by Schaid [1996] in the context of family data). These models, which are additive on a logit scale, correspond to multiplicative models on the odds scale and, therefore, to multiplicative models on an absolute risk scale, under an assumption of rare disease.

The advantage of specifying the test as a generalized linear model is that it is easy, in this framework, to test the additional effect of a locus, once the effects of other loci have already been accounted for. We do this by comparing the fit of a model that includes both the locus of interest and the other loci to that of a model where only the other loci are included. By restricting the parameter(s) corresponding to the locus of interest to main effects, we can we can perform a 1-df or 2-df test for the additional effect of any SNP even after the effects of other SNPs have already been accounted for. For example, to test the effect of a third (test) locus while accounting for the main effects at two primary loci, and allowing for dominance at the primary and test loci, we compare the fit of the model

$$\text{logit}(p) = \beta_0 + \sum_{i=1}^{6} \beta_i x_i$$

with

$$\text{logit}(p) = \beta_0 + \sum_{i=1}^{4} \beta_i x_i$$

(where the pairs $[x_1, x_2]$, $[x_3, x_4]$, and $[x_5, x_6]$ are the coded genotypes for the three loci). We may also allow

for dominance at the conditioning (primary) loci but not at the test locus by comparing

$$\text{logit}(p) = \beta_0 + \sum_{i=1}^{5} \beta_i x_i$$

with

$$\text{logit}(p) = \beta_0 + \sum_{i=1}^{4} \beta_i x_i \ .$$

To test the main effect of the test locus, while accounting for the full genotype effect of the two primary loci, we compare

$$\text{logit}(p) = \beta_0 + \sum_{i=1}^{6} \beta_i x_i + \sum_{i=1}^{2} \sum_{j=3}^{4} \beta_{ij} x_i x_j$$

with

$$\text{logit}(p) = \beta_0 + \sum_{i=1}^{4} \beta_i x_i + \sum_{i=1}^{2} \sum_{j=3}^{4} \beta_{ij} x_i x_j \ .$$

A convenient strategy for evaluation of the effects of the different polymorphisms within the gene is to fit these models in a stepwise logistic-regression procedure. Many standard statistics packages will perform automated stepwise logistic regression, using forward or backward selection. In forward selection, we start with the null model $\text{logit}(p) = \beta_0$ and, for each locus in turn, consider the effect of adding the locus to the model. Provided that at least one of the loci gives a significant improvement in fit, we add the most significant locus to the current model and then consider the effect of adding additional loci. Again, each additional locus is considered in turn, and the one giving the greatest improvement in fit is added to the current model. This procedure is repeated until no more loci give a significant improvement in fit. An alternative—and usually preferred—strategy is the use of backward selection, in which we start with a model that includes the effects at all polymorphisms under consideration. Each polymorphism is then deleted in turn from this full model, and the one that gives the least-significant deterioration in fit is removed from the current model. At the next stage, each locus still in the model is again deleted in turn, and the one that gives the least-significant deterioration is again removed from the current model. This continues until no more loci can be removed without significantly deteriorating the fit. In practice, most automated procedures employ a mixture of backward and forward selection, so that, at each stage, loci that have previously been removed may be added again if they offer significant improvement to the current model and vice versa. In this way, we generate a final model in which all polymorphisms in the model may be considered to have a significant effect on disease development, even once the effects of any other polymorphisms in that gene have been accounted for, and any polymorphisms not in the model do not appear to have significant effects once the polymorphisms that are in the model have been accounted for. The assumption is, therefore, that each of those polymorphisms in the model is likely to be either genuinely functional or to be in LD with another functional polymorphism that has not been typed in the original data set. Given a set of functional polymorphisms—which may be generated using the procedure above, from previous studies, or from knowledge about the functional variants in the region—we may test the overall effect of the set of variants by comparing the model that includes these variants to the null model in which none are included. However, if the set consists of a large number of polymorphisms, this test will have a large number of degrees of freedom and so may not be as powerful as individual tests of the individual polymorphisms.

Nuclear Family (TDT) Data

Suppose we have a sample of N unrelated cases with parents, all of whom are again genotyped at L polymorphisms (loci) within a gene that is believed to be associated with disease. If the haplotypes across the gene in the case and parents have been determined (i.e., phase-known data), the appropriate likelihood is the conditional logistic-regression likelihood (Self et al. 1991; Schaid 1996) in which each case is matched to three "pseudocontrols" such that the genotypes of the pseudocontrols consist of the other three possible haplotype combinations that could have been transmitted from the parents. Suppose, for example, that we have parents with haplotypes *a/b* and *c/d*, and the affected offspring has the haplotype *a/c*. Denote the fully phased genotypes of the case, the mother, and the father as g_c, g_m, and g_f, respectively, and let D represent the event that an individual is a case (i.e., diseased). Then,

$$P(g_c|g_m,g_f,D) = \frac{P(D|g_c,g_m,g_f)P(g_c|g_m,g_f)P(g_m,g_f)}{\sum_{g^* \in G} P(D|g^*,g_m,g_f)P(g^*|g_m,g_f)P(g_m,g_f)}$$

$$= \frac{P(D|g_c)P(g_c|g_m,g_f)}{\sum_{g^* \in G} P(D|g^*)P(g^*|g_m,g_f)} \ ,$$

where the sum in the denominator is over the four possible haplotype combinations that the parents can produce and each g^* is one of those four. Let $R(g)$ be the relative risk of disease for genotype g relative to some

arbitrarily chosen baseline genotype. Then this equation simplifies to

$$P(g_c|g_m,g_f,D) = \frac{R(g_c)}{\sum_{g^* \in G} R(g^*)} \quad (1)$$

or, in our example,

$$P(g_c|g_m,\ g_f,\ D) = \frac{R(a/c)}{R(a/c) + R(a/d) + R(b/c) + R(b/d)} \ . \quad (2)$$

As already noted, this likelihood is equivalent to that used in conditional logistic regression with a case with (phased) genotype *a/c* matched to three controls with (phased) genotypes *a/d*, *b/c*, and *b/d*. The likelihood for the whole data set is the product of these terms across all N trios of cases and their parents. This may be conveniently calculated by use of standard software for conditional logistic regression.

If the effect of the haplotypes is assumed to be multiplicative—that is, if $R(i/j)$ can be written in terms of haplotype relative risk parameters r_i and r_j, such that $R(i/j) = r_i r_j$—we have

$$P(g_c|g_m,g_f,D) = \frac{r_a r_c}{r_a r_c + r_a r_d + r_b r_c + r_b r_d} = \frac{r_a}{r_a + r_b} \cdot \frac{r_c}{r_c + r_d} \ , \quad (3)$$

so the parental transmissions can be considered as independent. Note that this is equivalent to assuming multiplicative effects of alleles within loci and also multiplicative effects of loci (Koeleman et al. 2000). This assumption will be true under the null hypothesis that there is no relationship between genotype (at any of the loci) and disease (i.e., $R(g) = 1, \forall g$) but may not hold under other, more complex null hypotheses—for example, the hypothesis that some of the loci, but not others, are involved in disease.

The problem with this approach in analysis of multilocus data is that, often, the haplotypes are not known (although they may, in some cases, be inferred). Therefore, the appropriate conditional probability expression for a trio is not $P(g_c|g_m,g_f,D)$ but instead $P(G_c|G_m,G_f,D)$, where G_i refers to the unphased genotypes of individual *i*. Analogously to the phase-known situation, we may write

$$P(G_c|G_m,G_f,D) = \frac{P(D|G_c)P(G_c|G_m,G_f)}{\sum_{G^* \in G} P(D|G^*)P(G^*|G_m,G_f)} \ . \quad (4)$$

The expressions $P(G_c|G_m,G_f)$ and $P(G^*|G_m,G_f)$ involve prior probabilities of parental phase and recombination probabilities between the loci. Specifically,

$$\begin{aligned} P(G^*|G_m,G_f) &= \sum_{g_m,g_f} P(G^*|g_m,g_f) \\ &\quad \times P(g_m,g_f|G_m,G_f) \\ &= \sum_{g_m,g_f} P(G^*|g_m,g_f) \\ &\quad \times \frac{P(G_m,G_f|g_m,g_f)P(g_m,g_f)}{P(G_m,G_f)} \ , \end{aligned}$$

where $P(g_m,g_f)$ and $P(G_m,G_f)$ are functions of the underlying population haplotype (or phased genotype) frequencies, $P(G_m,G_f|g_m,g_f)$ take values of 0 or 1, depending on whether G_m and G_f are compatible with g_m and g_f, and $P(G^*|g_m,g_f)$ depend on Mendelian transmission and recombination. Even if the recombination rate is assumed to be 0 (since the loci are within the same gene), the fact that G_m and G_f may result from several different possibilities for g_m and g_f means that $P(G^*|G_m,G_f)$ may be non-zero for a large number of values of G^*. For example, given parents who are heterozygous for different alleles at each of L loci, there are 4^L possible values of G^*. Therefore, the sum in the denominator of equation (4) is over not four but a potentially large number of possible haplotype transmissions from the parents, with probabilities dependent on the underlying population frequencies of these haplotypes. This means the likelihood no longer has the convenient formulation of equation (2). However, it is possible to recover a similar formulation by further conditioning. Let ξ denote some event in the family on which we will condition. Then

$$P(G_c|\xi,D) = \frac{P(D|G_c)P(G_c|\xi)}{\sum_{G^* \in G} P(D|G^*)P(G^*|\xi)} \quad (5)$$

We will consider four choices for ξ that lead to likelihoods that can be conveniently expressed in a conditional logistic-regression framework. In each of these, having observed the parental and offspring genotypes G_m, G_f, G_c, we first condition upon the parental genotypes G_m, G_f and the disease status of the offspring, D. In general, the probabilities of the possible offspring genotypes G^* depend on population haplotype frequencies and recombination, as well as on genotype relative risks. To avoid the dependence of the likelihood on these nuisance parameters, we condition further by identifying subsets of possible offspring genotypes, within which the probability of each offspring genotype, given parental genotypes, is constant. Then, conditioning upon the off-

spring genotype belonging to such a subset, say S, the probability $P(G_c|S,D)$ can be written as

$$P(G_c|S,D) = \frac{P(D|G_c)P(G_c|S)}{\sum_{G^* \in S} P(D|G^*)P(G^*|S)}$$

$$= \frac{P(D|G_c)}{\sum_{G^* \in S} P(D|G^*)} .$$

This is the same as the likelihood for a matched case/control study in which the case is the genotype of the affected offspring and the controls are the other genotypes in the set S. This likelihood therefore can be conveniently fitted using standard software for conditional logistic regression. Moreover, since the likelihood obtained is a true conditional likelihood, the parameter estimates obtained will have the properties of conditional maximum likelihood estimates—that is, they will provide consistent estimates of the underlying genotype (or, where appropriate, haplotype) relative risks.

Method 1. Conditioning on the fact that all possible offspring allow deduction of parental phase: the CETDT.—For the first conditioning method, we define subsets S in terms of parental phase. Given unphased parental genotypes G_m, G_f, there are a number of possible underlying phased genotypes g_m, g_f. We define S_k as the set of the four possible unphased offspring genotypes transmitted from the parents, assuming phase assignment k and no recombination between the loci. For instance, suppose the father has unphased genotypes a/b at locus 1 and u/v at locus 2, which we denote (a/b, u/v) and the mother has genotypes c/d at locus 1 and x/y at locus 2, denoted (c/d, x/y). The possible phases in the parents are: phase 1, au/bv and cx/dy; phase 2, au/bv and cy/dx; phase 3, av/bu and cx/dy; phase 4, av/bu and cy/dx. The resulting subsets of offspring genotypes are, therefore,

$$S_1 = \{(a/c, u/x), (a/d, u/y), (b/c, v/x), (b/d, v/y)\} ,$$

$$S_2 = \{(a/c, u/y), (a/d, u/x), (b/c, v/y), (b/d, v/x)\} ,$$

$$S_3 = \{(a/c, v/x), (a/d, v/y), (b/c, u/x), (b/d, u/y)\} ,$$

and

$$S_4 = \{(a/c, v/y), (a/d, v/x), (b/c, u/y), (b/d, u/x)\} .$$

Note that elements within a subset may be identical; for example, if the mother's genotype were also (a/b, u/v) instead of (c/d, x/y), then the first set would be $S_1 = \{(a/a, u/u), (a/b, u/v), (a/b, u/v), (b/b, v/v)\}$ which has only three distinct elements, since the unphased genotype (a/b, u/v) appears twice. Note that elements in different subsets may also be identical; for example, in the case where mother and father both have genotype (a/b, u/v), we also have $S_4 = \{(a/a, v/v), (a/b, u/v), (a/b, u/v), (b/b, u/u)\}$—that is, the genotype ($a/b$, u/v) appears in both S_1 and S_4.

Subsets constructed in this way have the property that the prior probability of all the elements within a subset (given G_m and G_f) is constant, since the elements are constructed by Mendelian inheritance from a particular set of phased parental genotypes g_m, g_f. Although we assumed zero recombination in the construction of the subsets, this assumption is not necessary to achieve the property of constant prior probability; for example, if the recombination fraction (θ) were allowed between the loci, given underlying parental phase 1, we would be able to construct the elements in subsets S_1, S_2, S_3, and S_4 with probabilities $(1-\theta)^2$, $\theta(1-\theta)$, $(1-\theta)\theta$, and θ^2, respectively. Since we will be conditioning on being in a particular subset, it does not matter which underlying parental phase gave rise to that subset; the recombination parameter is conditioned out in the likelihood. An equivalent way of considering this is to define the whole procedure as a function of gametic haplotypes—that is, those produced by the parents at meiosis (Koeleman et al. 2000)—rather than as a function of the actual parental haplotypes.

Conditional on the parental genotypes G_m, G_f, we construct phase-specific subsets $S_1, S_2, \ldots, S_P$, where P denotes the number of possible phase assignments in the parents. The observed offspring genotype G_c may fall into one or more of these subsets. Falling into one and only one subset implies that the parental phase can be inferred unambiguously from the offspring genotype. Given that G_c falls into a particular subset S_i, if that subset is disjoint from (i.e., has no elements in common with) all other subsets, then G_c must only appear in subset S_i. For method 1, we condition on the fact that $G_c \in S_i$ and that S_i is disjoint from all other subsets S_k. The overall conditioning event ξ, therefore, corresponds to the intersection of the events that the parental genotypes are G_m and G_f, that the offspring is affected, that the offspring genotype G_c is in the set S_i, and that $S_i \bigcap S_k = \{\}, \forall k \neq i$, with S_i itself only being determined after the offspring genotype G_c is observed. This conditioning is essentially equivalent to that used by the CETDT (Koeleman et al. 2000). Although the fact is not entirely clear from the original description of the CETDT, this test first deduces parental haplotypes and then uses only families in which the same parental haplotypes would be deducible from all possible offspring that parents with those deduced haplotypes could produce. (In this way, the bias that can occur when only unambiguous haplotype transmissions are counted [see Dudbridge et al. 2000] is avoided.) The resulting data consists of one case and three pseudocontrols that can be analyzed using conditional logistic regression. Suppose we deduce parental haplotypes a/b and c/d and

affected offspring haplotypes *a*/*c*. Then, provided that all possible offspring *a*/*c*, *a*/*d*, *b*/*c*, and *b*/*d* allow deduction of the same parental haplotypes, we use *a*/*c* as the case and *a*/*d*, *b*/*c*, and *b*/*d* as the three pseudocontrols. If deduction of parental phase is not possible from all pseudocontrols, we discard the family entirely (more precisely, the multiplicative likelihood contribution from this family is 1). For example, suppose we have a family that is typed at two SNPs, with the father and mother both heterozygous 1/2 at both loci and with the affected child homozygous 1/1 at both loci. Then we know from the child that, in each parent, the haplotypes must be 11/22. In this case, we can deduce the transmitted and untransmitted haplotypes: each parent transmits an 11 haplotype and does not transmit a 22 haplotype. However, if the offspring instead had received a 11 haplotype from one parent and a 22 from the other, we would not be able to tell whether the transmitted haplotypes were 11 and 22 or 12 and 21. The CETDT therefore discards this family, since the parental haplotypes are not deducible from all possible offspring. A way to avoid discarding the family would be to implement the CETDT, using an expectation-maximization (EM) algorithm (Dempster et al. 1977), which would allow all families to contribute to the test statistic, even when the parental phase is uncertain. However, implementation of a valid EM procedure when a complex null hypothesis is being tested is problematic. Although such fitting is not required by the theory, a major limitation of the CETDT as currently implemented is that it fits the conditional version of likelihood (3) rather than likelihood (2)—that is, a multiplicative model is assumed for the effects of different loci and for alleles within a single locus (Koeleman et al. 2000). This is necessary to achieve the factorization given in equation (3) and thus to consider maternal and paternal transmissions as independent. In the next section, we will consider likelihoods that do not have this restriction and that also, through a modification of the conditioning, allow utilization of data from a larger number of families.

Method 2. Conditioning on phase being inferable.—The fact that method 1 discards families in which the parental phase is not inferable from all possible offspring is likely to result in a loss of information, particularly when using extended haplotypes of loci that are not very polymorphic. By adjusting the conditioning argument, however, we may retain many of these families. In method 2 we proceed exactly as in method 1, except that, having determined a phase-specific subset S_i to which G_c belongs, we condition on G_c belonging only to the "disjoint" part of S_i—that is, the subset S_i' of S_i that does not intersect with any other phase-specific subset S_k. If S_i is disjoint from all S_k, the resulting subset S_i' on which we condition is S_i, the same as in method 1. If S_i has some elements in common with any S_k, S_i' consists of just those elements of S_i that do not appear in any other S_k. The distinction between this approach and that of method 1 is that if any of the possible offspring (in addition to the case) allow deduction of g_m and g_f, we use these individuals as pseudocontrols, together with the case, in a conditional logistic regression analysis. Suppose we deduce parental haplotypes *a*/*b* and *c*/*d* and affected offspring haplotypes *a*/*c*. Then, in this case, the denominator of (5) would sum over those possible offspring that allow deduction of parental phase—that is, the sum is over G^* corresponding to *a*/*c* (for the case), and between one and three additional values for G^*, corresponding to whichever of the offspring *a*/*d*, *b*/*c*, and *b*/*d* allow deduction of parental phase. In our earlier example—where both parents were heterozygous 1/2 at both loci, and the affected child was homozygous 1/1 at both loci—we therefore have a case with two 11 haplotypes and a single pseudocontrol with two 22 haplotypes, since these are the only possible offspring that would allow deduction of parental phase. The likelihood contribution for this family, therefore, would be

$$P(G_c|\xi,D) = \frac{R(G_c)}{\sum_{G^* \in G} R(G^*)} = \frac{R(11/11)}{R(11/11) + R(22/22)}$$

Note that use of this likelihood contribution is equivalent to that achieved when the correct variance is used in a TDT procedure with extended haplotypes (Dudbridge et al. 2000).

This approach extends very naturally to the situation where there is additional genetic information in a family—for example, genotypes of unaffected siblings or additional family members. We can use all available information to deduce parental phase. If phase is not deducible, this family will make no contribution to the likelihood. If phase is deducible, we consider the other three possible phased genotypes that the case could have inherited from the parents. These form three potential pseudocontrols. For each pseudocontrol, in turn, we test whether, if the index or case offspring were replaced by the pseudocontrol, the phased parental genotypes would still be uniquely determined (taking into account the additional family information—for example, from unaffected offspring). If the parental haplotypes are still inferable, the pseudocontrol is retained. Provided that there is at least one pseudocontrol remaining, we use the index case and matched pseudocontrol(s) in the conditional logistic-regression analysis.

Method 3. Conditioning on the set of transmitted and untransmitted genotypes, regardless of phase.—The two

methods described above are based on an ability to infer gametic phase. If one is interested only in whether a polymorphism is causal or not—that is, interested in fitting models on the basis of the main effects of the genotypes at L polymorphisms—one could argue that phase and, hence, haplotype determination are irrelevant. (This contrasts with linkage disequilibrium mapping of a single polymorphism, when one is indeed interested in determining an ancestral haplotype to localize the functional polymorphism.) The advantage of ignoring phase is that it allows us to use families that would be discarded in the previous two methods. In this case, we define the subsets S differently, not according to parental phase. Given G_m, G_f, and G_c, we define t_l to be the genotype consisting of the transmitted alleles and u_l to be the genotype consisting of the untransmitted alleles, at locus l. For the affected child, therefore, t_l is the unphased genotype at that locus, and u_l can be calculated as the genotype made up from the two parental alleles "left over" when the transmitted alleles are removed from the set of four parental alleles at that locus. We define vectors $\mathbf{t}$ and $\mathbf{u}$ as $\mathbf{t} = (t_1, t_2, \ldots, t_L)$ and $\mathbf{u} = (u_1, u_2, \ldots, u_L)$. We now define S, the set to be conditioned on, as $S = \{\mathbf{t}, \mathbf{u}\}$. A priori, $\mathbf{t}$ and $\mathbf{u}$ are equiprobable, because, if $\mathbf{t}$ occurs in any phase-specific subset S_i, $\mathbf{u}$ must also occur in S_i; thus, different probabilities of different phases cannot result in different prior probabilities for $\mathbf{t}$ and $\mathbf{u}$. With this conditioning, only two terms appear in the denominator of equation (5), corresponding to the (unphased) genotypes of the index case and a single pseudocontrol whose unphased genotypes at each locus are made up from the untransmitted alleles from each parent. For example, suppose that a family is typed at three SNPs, with the father having genotypes (1/1, 1/2, and 1/2) at loci (1, 2, and 3), respectively; the mother having genotypes (1/2, 1/2, and 1/2); and the affected offspring having genotypes (1/1, 1/2, and 1/2). This family would not be used in methods 1 or 2 above, since the parental haplotypes cannot be determined. Ignoring phase, however, we have a single case with genotypes (1/1, 1/2, and 1/2) and a single matched pseudocontrol with genotypes (1/2, 1/2, and 1/2). Writing the genotype relative risks for genotype (i, j, and k) at loci (1, 2, and 3) as $R(i, j, k)$, we therefore have a likelihood contribution, for this family, of

$$\frac{R(1/1,\ 1/2,\ 1/2)}{R(1/1,\ 1/2,\ 1/2) + R(1/2,\ 1/2,\ 1/2)}\ .$$

Note that this likelihood will only be valid for fitting models in which relative risk does not depend on phase.

Method 4. Conditioning according to method 2 when phase is deducible and to method 3 otherwise.—Use of a single, matched pseudocontrol instead of as many as three matched pseudocontrols might be expected to reduce efficiency for families where parental haplotypes can, in fact, be determined. For instance, if a single locus is being considered, with parental genotypes a/b, c/d, and affected offspring genotype a/c, method 3 uses a single pseudocontrol with genotype b/d and ignores the a/d and b/c pseudocontrols. A maximally efficient set of cases and pseudocontrols can be obtained by use of conditioning method 4, which consists of using the strategy of method 2 when parental phase is deducible and method 3 otherwise. In this case, the method for constructing conditioning sets S will differ from family to family, being determined only after observation of G_m, G_f, G_c. This strategy allows every family to be used, generating one case and between one and three pseudocontrols for each family, regardless of whether parental phase can be determined. Again, this method will only be valid for fitting models in which relative risk does not depend on phase.

The relative efficiencies of methods 3 and 4 can be calculated analytically in simple situations (from consideration of the relative variances of the parameter estimates, as calculated from the expected information matrix). For example, figure 1 shows the relative efficiency of method 4, compared with method 3, when a single diallelic locus is considered. The relative efficiency depends on the allele frequency and haplotype relative risk associated with allele 1. Note that, as the relative risk approaches 1, the relative efficiency also approaches 1, so there is no loss of efficiency with method 3 when testing the null hypothesis of no effect. For higher values of the haplotype relative risk, however, depending on the allele frequencies, method 4 can be substantially

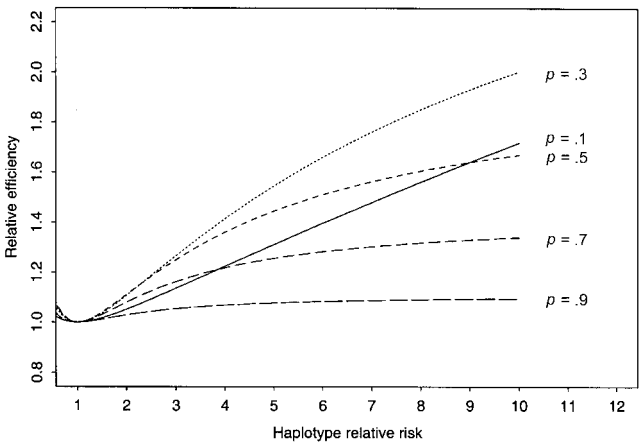

Figure 1 Relative efficiency of method 4, compared with that of method 3, for a single diallelic locus. Results are given as a function of haplotype relative risk due to allele 1, where the population frequency of allele 1 is p.

more efficient than method 3, suggesting that in these cases there will be an advantage in using method 4.

Stepwise Regression Procedure

We evaluate the effects of the individual polymorphisms using exactly the same stepwise approach as described for case/control data. We write $P(D|g)$ or $P(D|G)$ as being proportional to genotype relative risks $R(g)$ or $R(G)$. A convenient parameterization (Schaid 1996) is

$$R(g) = e^{\beta^T x},$$

where $\mathbf{x}$ is the coded genotype for the case or pseudocontrol. Note that $\beta_0 = 0$ in this parameterization, since risks are all evaluated relative to an arbitrary reference genotype. If method 1 is used and we are willing to assume a multiplicative model, we may instead code according to haplotypes instead of genotypes, using haplotypes or "chromosomes" transmitted by the two parents as independent data points. In either case, by considering the fit of a series of nested models, we can again evaluate the effect of individual polymorphisms once other polymorphisms have already been accounted for.

The CETDT proposed by Koeleman et al. (2000) uses a similar procedure to evaluate the effect of a polymorphism while taking into account effects at other loci. However, the CETDT compares the fit of models where all genotype effects (or haplotype effects, under a multiplicative model) are estimated, instead of fitting only the main effects of loci. This results in a test with a much larger number of degrees of freedom than the tests proposed here. Consequently, the procedure described here is likely to have much higher power for evaluation of the main effects of polymorphisms. We illustrate the difference between the parameterizations in table 1, under the assumption that we are interested in evaluating the effect of a third SNP while taking into account effects at two other SNPs. For the genotype test, the CETDT would test a model with 26 parameters against a null model with 8 parameters, resulting in an 18-df test. This contrasts with the proposed regression procedure, which tests a model with six parameters against a null model with four parameters (allowing for dominance), resulting in a 2-df test. Note that, if desired, the full 26-parameter model or any intermediate models could also be fitted in the regression procedure by including statistical interaction terms in the model. For the haplotype test, the CETDT would compare models with seven and three parameters, resulting in a 4-df test, whereas the proposed regression procedure compares models with three and two parameters, resulting in a 1-df test. The parameterization shown in table 1 illustrates that, even when testing haplotypes, the proposed coding for the regression method results in likelihood contributions for the case, given its (unphased) genotype, that are identical for the different possible haplotypes in the parents. This reflects the fact that the coding system models the probability of being affected with disease only as a function of the (unphased) genotypes. Likelihood contributions for the pseudocontrols will, however, depend on the parental haplotypes (through the resulting pseudocontrol genotypes), except for regression method 3, in which only the single pseudocontrol with genotypes constructed from the untransmitted parental alleles is used.

Sibships with Multiple Affected Sibs

The approaches described here for case/control and family-based analyses can all be extended to situations where there is more than one affected offspring per pedigree. These offspring cannot be used as independent observations in a test of association, because there is likely to be linkage to disease in the region (i.e., the same alleles are likely to be transmitted to any affected offspring). Martin et al. (1997) have overcome this problem by devising a test for association that uses data from all affected offspring in a sibship. A natural way to implement this in the context of the regression methods described here is to test the fit of nested hypotheses using a Wald test rather than a likelihood-ratio test. A robust "information-sandwich" estimate of the variance/covariance matrix for β (Huber 1967; White 1982) may be obtained, which allows us to perform the Wald test while accounting for the nonindependence between affected siblings in the same family.

Missing Data and Ambiguous Haplotypes

Use of standard statistical packages for stepwise logistic and conditional logistic-regression forces some limitations in the treatment of missing data. Most packages will only include in the analysis individuals (i.e., cases, controls, or pseudocontrols) who are fully genotyped at every locus that can potentially contribute to the current model. If large numbers of polymorphisms are to be considered for inclusion—even if the rate of missing data at each locus is low—the proportion of individuals who have missing genotypes at one or more loci and therefore must be discarded from the analysis can be quite high. If different sample sizes are used for different model comparisons, to use the maximal amount of information available for that comparison, the powers of the different steps in the model-building procedure may vary. This would make it difficult to truly compare the significances of adding or deleting terms at different stages. For this reason, in our *HLA* analyses

Table 1

Parameterization for Genotype Relative Risks *R*(*g*) Used by Different Methods

Case Genotype at Loci 1, 2, 3	Resulting Haplotypes	Genotype Test				Haplotype Test			
		Regression Method		CETDT		Regression Method		CETDT	
		Null	Alt	Null	Alt	Null	Alt	Null	Alt
1/1, 1/1, 1/1	111/111	1	1	1	1	1	1	1	1
1/1, 1/1, 1/2	111/112	1	e^{β_5}	1	R_{111112}	1	e^{β_3}	1	$1 \cdot R_{112}$
1/1, 1/1, 2/2	112/112	1	e^{β_6}	1	R_{111122}	1	$e^{\beta_3} \cdot e^{\beta_3}$	1	$R_{112} \cdot R_{112}$
1/1, 1/2, 1/1	111/121	e^{β_3}	e^{β_3}	R_{1112}	R_{111211}	e^{β_2}	e^{β_2}	R_{12}	$1 \cdot R_{121}$
1/1, 1/2, 1/2	111/122 or 112/121	e^{β_3}	$e^{\beta_3+\beta_5}$	R_{1112}	R_{111212}	e^{β_2}	$e^{\beta_2+\beta_3}$	R_{12}	$1 \cdot R_{122}$ or $R_{112} \cdot R_{121}$[a]
1/1, 1/2, 2/2	112/122	e^{β_3}	$e^{\beta_3+\beta_6}$	R_{1112}	R_{111222}	e^{β_2}	$e^{\beta_3}e^{\beta_2+\beta_3}$	R_{12}	$R_{112} \cdot R_{122}$
1/1, 2/2, 1/1	121/121	e^{β_4}	e^{β_4}	R_{1122}	R_{112211}	$e^{\beta_2}e^{\beta_2}$	$e^{\beta_2}e^{\beta_2}$	$R_{12} \cdot R_{12}$	$R_{121} \cdot R_{121}$
1/1, 2/2, 1/2	121/122	e^{β_4}	$e^{\beta_4+\beta_5}$	R_{1122}	R_{112212}	$e^{\beta_2}e^{\beta_2}$	$e^{\beta_2}e^{\beta_2+\beta_3}$	$R_{12} \cdot R_{12}$	$R_{121} \cdot R_{122}$
1/1, 2/2, 2/2	122/122	e^{β_4}	$e^{\beta_4+\beta_6}$	R_{1122}	R_{112222}	$e^{\beta_2}e^{\beta_2}$	$e^{\beta_2+\beta_3}e^{\beta_2+\beta_3}$	$R_{12} \cdot R_{12}$	$R_{122} \cdot R_{122}$
1/2, 1/1, 1/1	111/211	e^{β_1}	e^{β_1}	R_{1211}	R_{121111}	e^{β_1}	e^{β_1}	R_{21}	$1 \cdot R_{211}$
1/2, 1/1, 1/2	111/212 or 112/211	e^{β_1}	$e^{\beta_1+\beta_5}$	R_{1211}	R_{121112}	e^{β_1}	$e^{\beta_1+\beta_3}$	R_{21}	$1 \cdot R_{212}$ or $R_{112} \cdot R_{211}$[a]
1/2, 1/1, 2/2	112/212	e^{β_1}	$e^{\beta_1+\beta_6}$	R_{1211}	R_{121122}	e^{β_1}	$e^{\beta_3}e^{\beta_1+\beta_3}$	R_{21}	$R_{112} \cdot R_{212}$
1/2, 1/2, 1/1	111/221 or 121/211	$e^{\beta_1+\beta_3}$	$e^{\beta_1+\beta_3}$	R_{1212}	R_{121211}	$e^{\beta_1+\beta_2}$	$e^{\beta_1+\beta_2}$	R_{22} or $R_{12} \cdot R_{21}$	$1 \cdot R_{221}$ or $R_{121} \cdot R_{211}$[a]
1/2, 1/2, 1/2	111/222 or 112/221 or 121/212 or 211/122	$e^{\beta_1+\beta_3}$	$e^{\beta_1+\beta_3+\beta_5}$	R_{1212}	R_{121212}	$e^{\beta_1+\beta_2}$	$e^{\beta_1+\beta_2+\beta_3}$	R_{22} or $R_{12} \cdot R_{21}$	$1 \cdot R_{222}$ or $R_{112} \cdot R_{221}$ or $R_{121} \cdot R_{212}$ or $R_{211} \cdot R_{122}$[a]
1/2, 1/2, 2/2	112/222 or 122/212	$e^{\beta_1+\beta_3}$	$e^{\beta_1+\beta_3+\beta_6}$	R_{1212}	R_{121222}	$e^{\beta_1+\beta_2}$	$e^{\beta_1+\beta_2+2\beta_3}$	R_{22} or $R_{12} \cdot R_{21}$	$R_{112} \cdot R_{222}$ or $R_{122} \cdot R_{212}$[a]
1/2, 2/2, 1/1	121/221	$e^{\beta_1+\beta_4}$	$e^{\beta_1+\beta_4}$	R_{1222}	R_{122211}	$e^{\beta_2}e^{\beta_1+\beta_2}$	$e^{\beta_2}e^{\beta_1+\beta_2}$	$R_{12} \cdot R_{22}$	$R_{121} \cdot R_{221}$
1/2, 2/2, 1/2	121/222 or 122/221	$e^{\beta_1+\beta_4}$	$e^{\beta_1+\beta_4+\beta_5}$	R_{1222}	R_{122212}	$e^{\beta_1+2\beta_2}$	$e^{\beta_1+2\beta_2+\beta_3}$	$R_{12} \cdot R_{22}$	$R_{121} \cdot R_{222}$ or $R_{122} \cdot R_{221}$[a]
1/2, 2/2, 2/2	122/222	$e^{\beta_1+\beta_4}$	$e^{\beta_1+\beta_4+\beta_6}$	R_{1222}	R_{122222}	$e^{\beta_2}e^{\beta_1+\beta_2}$	$e^{\beta_2+\beta_3}e^{\beta_1+\beta_2+\beta_3}$	$R_{12} \cdot R_{22}$	$R_{122} \cdot R_{222}$
2/2, 1/1, 1/1	211/211	e^{β_2}	e^{β_2}	R_{2211}	R_{221111}	$e^{\beta_1}e^{\beta_1}$	$e^{\beta_1}e^{\beta_1}$	$R_{21} \cdot R_{21}$	$R_{211} \cdot R_{211}$
2/2, 1/1, 1/2	211/212	e^{β_2}	$e^{\beta_2+\beta_5}$	R_{2211}	R_{221112}	$e^{\beta_1}e^{\beta_1}$	$e^{\beta_1}e^{\beta_1+\beta_3}$	$R_{21} \cdot R_{21}$	$R_{211} \cdot R_{212}$
2/2, 1/1, 2/2	212/212	e^{β_2}	$e^{\beta_2+\beta_6}$	R_{2211}	R_{221122}	$e^{\beta_1}e^{\beta_1}$	$e^{\beta_1+\beta_3}e^{\beta_1+\beta_3}$	$R_{21} \cdot R_{21}$	$R_{212} \cdot R_{212}$
2/2, 1/2, 1/1	211/221	$e^{\beta_2+\beta_3}$	$e^{\beta_2+\beta_3}$	R_{2212}	R_{221211}	$e^{\beta_1}e^{\beta_1+\beta_2}$	$e^{\beta_1}e^{\beta_1+\beta_2}$	$R_{21} \cdot R_{22}$	$R_{211} \cdot R_{221}$
2/2, 1/2, 1/2	211/222 or 212/221	$e^{\beta_2+\beta_3}$	$e^{\beta_2+\beta_3+\beta_5}$	R_{2212}	R_{221212}	$e^{2\beta_1+\beta_2}$	$e^{2\beta_1+\beta_2+\beta_3}$	$R_{21} \cdot R_{22}$	$R_{211} \cdot R_{222}$ or $R_{212} \cdot R_{221}$[a]
2/2, 1/2, 2/2	212/222	$e^{\beta_2+\beta_3}$	$e^{\beta_2+\beta_3+\beta_6}$	R_{2212}	R_{221222}	$e^{\beta_1}e^{\beta_1+\beta_2}$	$e^{\beta_1+\beta_3}e^{\beta_1+\beta_2+\beta_3}$	$R_{21} \cdot R_{22}$	$R_{212} \cdot R_{222}$
2/2, 2/2, 1/1	221/221	$e^{\beta_2+\beta_4}$	$e^{\beta_2+\beta_4}$	R_{2222}	R_{222211}	$e^{\beta_1+\beta_2}e^{\beta_1+\beta_2}$	$e^{\beta_1+\beta_2}e^{\beta_1+\beta_2}$	$R_{22} \cdot R_{22}$	$R_{221} \cdot R_{221}$
2/2, 2/2, 1/2	221/222	$e^{\beta_2+\beta_4}$	$e^{\beta_2+\beta_4+\beta_5}$	R_{2222}	R_{222212}	$e^{\beta_1+\beta_2}e^{\beta_1+\beta_2}$	$e^{\beta_1+\beta_2}e^{\beta_1+\beta_2+\beta_3}$	$R_{22} \cdot R_{22}$	$R_{221} \cdot R_{222}$
2/2, 2/2, 2/2	222/222	$e^{\beta_2+\beta_4}$	$e^{\beta_2+\beta_4+\beta_6}$	R_{2222}	R_{222222}	$e^{\beta_1+\beta_2}e^{\beta_1+\beta_2}$	$e^{\beta_1+\beta_2+\beta_3}e^{\beta_1+\beta_2+\beta_3}$	$R_{22} \cdot R_{22}$	$R_{222} \cdot R_{222}$

NOTE.—Test is of whether locus 3 is involved once loci 1 and 2 have been accounted for. "Alt" = alternative model.

[a] Depends on which haplotypes are inferred in the parents.

Table 2

Size of Data Sets Generated Using Different Methods

	Multiallelic Coding			Diallelic Coding		
Method	Cases	Pseudocontrols	Total	Cases	Pseudocontrols	Total
1	156	468	624	112	336	448
2	209	521	730	190	414	604
3	247	247	494	247	247	494
4	247	559	806	247	471	718
CETDT[a]	312	312	624	224	224	448

[a] CETDT sample sizes refer to chromosomes rather than individuals.

below, we limit families to those that are fully typed at all five loci of interest or those in which both parents are homozygous for any locus at which the child is untyped. In practice, missing genotypes could be inferred and the uncertainty of such inference addressed by multiple imputation (Schafer 1997), the high level of LD within a gene ensuring that there will be little variation between imputations. A reasonable (although more complicated) strategy, therefore, might be to use multiple imputation to generate imputed complete data sets and then to analyze these in a standard statistical-analysis package.

A related problem occurs with family data when methods 1 and 2 are used. In this case, since phase is more easily resolved in haplotypes consisting of a few loci than in extended haplotypes containing many loci, to generate the maximum information for a particular test in the stepwise procedure, one would have to generate a different set of cases and pseudocontrols, depending on which set of markers are currently being tested. However, this is somewhat tortuous and, moreover, results in a different-sized data set and, therefore, in a different power to select or reject a locus depending on which stage in the stepwise-regression procedure we have reached. Note that this problem does not occur when method 3 is used, since no attempt is made to infer haplotypes or to resolve phase. In our *HLA* analyses below, we decided to use all five loci of interest to infer parental phase and used this data set even when analyzing just a subset of the loci. However, had we started with a much larger number of loci of interest, this approach would not have been satisfactory, since families in which phase could be resolved for a subset of the loci, but not for the whole set of loci, would not be used, even for analysis with the subset. In this case, use of multiple imputation methods, this time for inferring the fully phased genotypes of individuals, might again be suggested.

Results

The methods described here were used to evaluate the contributions of five polymorphisms in the *HLA* region to type 1 diabetes, using family data consisting of 353 trios of parents with a single affected offspring. These data were previously analyzed by Zavattari et al. (2001), using a variation of the haplotype method. (Note that the analysis by Zavattari et al. [2001] included an additional 32 multiplex families, with two affected children, that are not included here.) Families were typed at the loci *HLA-DRB1, HLA-DQB1, DMB, DOB,* and *TNFc.* For a family to be included in the analysis, we required that it be fully typed at all five loci of interest or that both parents be homozygous at any locus at which the child was untyped. This resulted in a final data set consisting of 247 trios, all of which could be used for methods 3 and 4. For methods 1 and 2, only those trios in which the five-locus parental haplotypes could be determined unambiguously could be used, with method 2 being further restricted to those trios in which at least one pseudocontrol would resolve phase, and method 1 even further restricted to those trios in which all pseudocontrols would resolve phase. Analysis was carried out using software written by the authors (see our Web site) for implementation within the statistical package Stata. The sizes of the resulting data sets available for the regression analysis are shown in columns 2–4 of table 2. Since the loci *DRB1, DQB1,* and *DMB* are relatively polymorphic, it was possible to resolve phase in a fairly large proportion of families, although not necessarily with all pseudocontrols. The final numbers of families retained when methods 1 and 2 were used were 156 and 209, respectively, compared to 247 for methods 3 and 4. This represents a minor loss of data when method 2 is used but a fairly severe loss with method 1. Methods 1 and 2 do, however, produce a larger number of pseudocontrols than method 3 (since each case in method 3 is matched to exactly one pseudocontrol, whereas, in method 1 and in some families for method 2, each case is matched to three pseudocontrols). Method 4 generates the largest data set, consisting of exactly one case and between one and three pseudocontrols from every family.

Although haplotypes were constructed using the full multiallelic coding for loci *DRB1, DQB1,* and *DMB,* for the regression analysis, the alleles were recoded in

a binary fashion in order to evaluate effects of a single allele of interest compared to all other alleles. Following Zavattari et al. (2001), the alleles examined were *DRB1*0301*, *DQB1*0201*, and *DMB*0101*. *DOB* and *TNFc* were already effectively diallelic with alleles 1 and 2 at *DOB* and 167 and 169 at *TNFc*. If recoding were carried out prior to haplotype determination, it would decrease the phase information and, thus, the number of cases and pseudocontrols produced by methods 1 and 2. Columns 5–7 of table 2 show the resulting sample sizes if binary recoding is carried out prior to haplotype determination. These data show a marked decrease in sample size for methods 1 and 2. It is clearly preferable to use a multiallelic coding at the haplotype determination stage. However, depending on the markers available, this may not be possible. Table 2, therefore, gives an indication of the loss of information to be expected for methods 1 and 2 when diallelic markers such as SNPs are used.

The results of various tests in the stepwise regression procedure are given in tables 3, 4, 5, and 6. For convenience, we have denoted the loci *DRB1*, *DQB1*, *DMB*, *DOB*, and *TNFc* as A, B, C, D, and E, respectively. Results are shown for methods 1–4, fitting full genotype (i.e., including dominance) but no haplotype effects at the loci being considered. The notation used is standard statistical modeling notation—so that, for example, "A" denotes a 2-df model with main effects at locus A only; "A+B" denotes a 4-df model with main effects at loci A and B only; "A*B" denotes an 8-df model with main effects at A and B, plus statistical interactions between them (i.e., accounting for the "full genotype effect" of A and B but no haplotype effects); and "A*B+C" denotes a 10-df model with main effects and statistical interactions at A and B and also main effects at C. Also shown in table 5 are results for method 1, under the assumption of a multiplicative model for loci and alleles, so that the factorization of equation (3) applies. For the tests in table 5, use of this procedure is equivalent to use of the CETDT. Overall, we find that methods 1–4 give broadly similar results, except that *P* values generally increase in significance across methods 1–4, as expected with the increase in sample size of the ensuing case/pseudocontrol data sets. Results from the CETDT are also similar—but recall that, for this procedure to be valid, multiplicative effects must be assumed. Since method 4 is based on the largest number of individuals, we use this as our reference method while noting where the results differ substantially from those obtained by use of methods 1–3.

Table 3 shows the results of testing the main effects of a locus in a forward stepwise regression procedure. At the first stage (rows 1–5) we see that all loci except C (*DMB*) appear to be significant when included in the model on their own i.e. without accounting for effects at any other loci, equivalent to carrying out a single-locus analysis. Once A (*DRB1*) is included (rows 6–9), only D (*DOB*) has a significant main effect. Once A and D are included as main effects (rows 10–12) only method 3 suggests borderline significant main effects at

Table 3

Likelihood-Ratio Tests of Main Effects in a Forward Stepwise-Regression Procedure

Null Model	Alternative Model	Method 1			Method 2			Method 3			Method 4		
		χ^2	df	*P*	χ^2	df	*P*	χ^2	df	*P*	χ^2	df	*P*
…	A	76.89	2	0[a]	117.10	2	0[a]	120.09	2	0[a]	124.92	2	0[a]
…	B	62.21	2	0[a]	105.73	2	0[a]	98.86	2	0[a]	101.06	2	0[a]
…	C	.05	2	.97	1.12	2	.57	1.53	2	.47	1.96	2	.37
…	D	9.72	2	.008	8.57	2	.01	9.62	2	.008	9.62	2	.008
…	E	15.79	2	.0004	44.53	2	2×10^{-10}	38.79	2	4×10^{-9}	38.79	2	4×10^{-9}
A	A+B	.86	2	.65	2.67	2	.26	5.68	2	.06	1.98	2	.37
A	A+C	3.61	2	.16	3.26	2	.20	2.05	2	.36	3.49	2	.17
A	A+D	13.33	2	.001	14.55	2	.0007	11.25	2	.004	17.36	2	.0002
A	A+E	.34	2	.85	2.11	2	.35	3.16	2	.21	1.88	2	.39
A+D	A+D+B	1.68	2	.43	3.10	2	.21	5.66	2	.06	2.56	2	.28
A+D	A+D+C	1.31	2	.52	.59	2	.75	.53	2	.77	.76	2	.69
A+D	A+D+E	.92	2	.63	2.21	2	.33	3.09	2	.21	1.99	2	.37
A*D	A*D+B	2.21	2	.33	2.79	2	.25	6.75	2	.03	3.66	2	.16
A*D	A*D+C	.65	2	.72	.60	2	.74	.32	2	.85	.61	2	.74
A*D	A*D+E	1.10	2	.58	2.52	2	.28	3.55	2	.17	2.51	2	.28
A+B+D	A+B+D+C	1.14	2	.57	.42	2	.81	.27	2	.87	.59	2	.74
A+B+D	A+B+D+E	.83	2	.66	2.36	2	.31	2.54	2	.28	1.64	2	.44
A*B*D	A*B*D+C	.58	2	.75	.34	2	.84	.03	2	.99	.23	2	.89
A*B*D	A*B*D+E	.73	2	.69	2.02	2	.36	2.62	2	.27	1.57	2	.46

NOTE.—A = *DRB*1, B = *DQB*1, C = *DMB*, D = *DOB*, and E = *TNFc*.

[a] $< 1 \times 10^{-10}$.

Table 4

Likelihood-Ratio Tests of Main Effects in a Backward Stepwise Regression Procedure

Null Model	Alternative Model	Method 1			Method 2			Method 3			Method 4		
		χ^2	df	P	χ^2	df	P	χ^2	df	P	χ^2	df	P
B+C+D+E	A+B+C+D+E	15.55	2	.0004	11.68	2	.003	20.63	2	3×10^{-5}	23.31	2	9×10^{-6}
A+C+D+E	A+B+C+D+E	1.42	2	.49	3.12	2	.21	4.98	2	.08	2.11	2	.35
A+B+D+E	A+B+C+D+E	1.15	2	.56	.24	2	.87	.18	2	.91	.43	2	.81
A+B+C+E	A+B+C+D+E	12.41	2	.002	12.52	2	.002	9.82	2	.007	15.49	2	.0004
A+B+C+D	A+B+C+D+E	.84	2	.66	2.18	2	.34	2.45	2	.29	1.48	2	.48
B*C*D*E	B*C*D*E+A	12.30	2	.002	10.74	2	.005	14.03	2	.0009	25.37	2	3×10^{-6}
A*C*D*E	A*C*D*E+B	.66	2	.72	2.01	2	.37	2.19	2	.34	1.47	2	.48
A*B*D*E	A*B*D*E+C	.43	2	.81	.88	2	.65	.25	2	.88	.42	2	.81
A*B*C*E	A*B*C*E+D	12.02	1	.0005	11.14	2	.004	9.83	2	.007	13.22	2	.001
A*B*C*D	A*B*C*D+E	1.40	2	.50	2.08	2	.35	2.48	2	.29	.93	2	.63
B+D	A+B+D	18.57	2	.0001	16.60	2	.0002	29.50	2	4×10^{-7}	29.27	2	4×10^{-7}
A+D	A+B+D	1.68	2	.43	3.10	2	.21	5.66	2	.06	2.56	2	.28
A+B	A+B+D	14.16	2	.0008	14.97	2	.0006	11.23	2	.004	17.94	2	.0001
D	A+D	80.50	2	0[a]	123.08	2	0[a]	121.73	2	0[a]	132.66	2	0[a]
A	A+D	13.33	2	.001	14.55	2	.0007	11.25	2	.004	17.36	2	.0002

NOTE.—A = *DRB1*, B = *DQB1*, C = *DMB*, D = *DOB*, and E = *TNFc*.

[a] $< 1 \times 10^{-10}$.

B (*DQB1*) (A + D + B vs. A + D; $P = .06$). If the full (main and interaction) effects of A and D are included (rows 13–15), again, only method 3 suggests borderline significant main effects at B (A*D+B vs. A*D, $P = .03$). Once A, D, and B have all been included in the model, no other loci show significant main effects (rows 16–19). This conclusion is supported by a backwards stepwise procedure (selected steps shown in table 4). Only loci A and D are significant when the main effects at all other loci have been included (rows 1–5; $P = 9 \times 10^{-6}$ for locus A, $P = .0004$ for locus D). Also, only loci A and D are significant when the full effect of all other loci has been included (rows 6–10; $P = 3 \times 10^{-6}$ for locus A, $P = .001$ for locus D). Continuing with the backwards stepwise procedure, the final model from method 4 is one with main effects at A (*DRB1*) and D (*DOB*) only. Neither locus can be deleted from this model without significantly worsening the fit (final two rows of table 4). With method 3, the final model could include locus B (*DQB1*) with borderline significance (rows 11–13, $P = 4 \times 10^{-7}$, $P = .06$, and $P = .004$, for worsening the fit when removing loci A, B, and D, respectively). Note that locus B (*DQB1*) along with *DRB1* (locus A) is believed to be a major determinant of the *IDDM1/HLA* locus in type 1 diabetes (Cucca et al. 2001*b*). However, the relative contributions of *DQB1* and *DRB1* depend on the relative frequencies of alleles and haplotypes at these loci in each specific population. Sardinia has a unique distribution of *DRB1*04* subtype alleles and the highest population frequency of *DR3* in the world, giving rise to a reduced relative contribution of the B (*DQB1*) locus in this population.

After main effects have been detected at A (*DRB1*), D (*DOB*), and, possibly, B (*DQB1*), it is of interest to fit interaction models to examine the full effect (i.e., main effect and/or statistical interactions with loci already included in the model) of the various loci (table 5). Once A has been included in the model, there is evidence for an effect at C and/or D ($P = .02$ for A*C vs. A; $P = 2 \times 10^{-7}$ for A*D vs. A). Once A and D have both been included, there is evidence for an effect at C (A*D*C vs. A*D; $P = .008$) and possibly at B with method 3 (A*D*B vs. A*D; $P = .04$). Once A, B, and D are included, there is still evidence for an effect at C (A*B*C*D vs. A*B*D; $P = .007$). Interestingly, with method 3, from the final five rows of table 5, we find that the full model A*B*C*D*E fits significantly better than any of the nested four-locus interaction models, suggesting that each of the loci makes a contribution to disease even when accounting for the full genotype contribution of the other four loci ($P = 3 \times 10^{-5}$, $P = .02$, $P = .0006$, $P = .0004$, and $P = .01$ for loci A–E, respectively. (With method 4, the evidence for the contribution of E is not significant, possibly because of the greater number of degrees of freedom.) These results should be interpreted with caution, because of the large number of estimated parameters and degrees of freedom for each test. This problem could potentially be overcome by use of a permutation procedure (Hirji et al. 1987, 1988). However, as presented, the results are consistent with those of Zavattari et al. (2001), who also showed that each of these five loci had a significant effect when considered on a particular background of alleles at two or three other loci. The difference here is that we are considering the contribution conditional on effects at all four other loci and regardless of which particular alleles are present at the loci.

Table 5

Likelihood-Ratio Combined Tests of Main Effects plus Interactions in a Stepwise-Regression Procedure

Null Model	Alternative Model	Method 1			Method 2			Method 3			Method 4			CETDT		
		χ^2	df	*P*	χ^2	df	*P*	χ^2	df	*P*	χ^2	df	*P*	χ^2	df	*P*
A	A*B	6.40	4	.17	7.39	4	.12	5.74	3	.13	5.62	4	.23	7.22	2	.03
A	A*C	11.16	6	.08	12.09	6	.06	10.79	6	.10	14.83	6	.02	13.60	2	.001
A	A*D	20.55	4	.0004	33.46	5	3×10^{-6}	28.12	5	3×10^{-5}	38.95	5	2×10^{-7}	24.45	2	5×10^{-6}
A	A*E	3.05	6	.80	6.27	6	.39	9.46	6	.15	5.28	6	.51	1.79	2	.41
A*D	A*D*B	10.19	6	.12	10.98	6	.09	11.86	5	.04	10.84	6	.09	14.00	3	.003
A*D	A*D*C	15.43	11	.16	18.59	13	.14	28.56	13	.008	28.42	13	.008	12.42	3	.006
A*D	A*D*E	11.88	12	.46	11.46	14	.65	11.07	13	.60	11.01	14	.69	3.15	4	.53
A*B*D	A*B*C*D	21.28	16	.17	26.18	18	.10	29.00	16	.02	36.13	18	.007	9.29	4	.05
A*B*D	A*B*D*E	25.08	20	.20	24.47	22	.32	22.81	19	.25	30.96	23	.12	5.66	7	.58
B*C*D*E	A*B*C*D*E	32.12	19	.03	38.73	19	.005	42.17	12	3×10^{-5}	61.21	21	8×10^{-6}	10.04	8	.26
A*C*D*E	A*B*C*D*E	21.05	20	.39	26.66	20	.15	25.26	13	.02	36.64	22	.03	10.65	8	.22
A*B*D*E	A*B*C*D*E	36.04	30	.21	51.34	32	.02	55.98	26	.0006	55.24	34	.01	12.90	8	.12
A*B*C*E	A*B*C*D*E	38.90	26	.05	54.11	31	.006	53.92	24	.0004	68.07	31	.0001	28.32	8	.0004
A*B*C*D	A*B*C*D*E	39.84	34	.23	49.62	36	.06	49.80	29	.01	50.07	39	.11	9.28	11	.60

NOTE.—A = *DRB1*, B = *DQB1*, C = *DMB*, D = *DOB*, and E = *TNFc*.

The fact that loci C (*DMB*) and E (*TNFc*) appear to be significant when tested in interaction with the other loci, but not when tested as main effects, suggests that it is loci A (*DRB1*), D (*DOB*) and, possibly, B (*DQB1*) that have the primary etiological effects in this Sardinian population, whereas C and E have modifying or interaction effects on the A-B-D contribution. This interpretation assumes that all functional polymorphisms in the region are available for inclusion in the model (i.e., all would be included in the first stage of a backwards stepwise procedure). If there are functional polymorphisms in the region not included in the procedure, detection of the main effect at A, for instance, may in fact correspond to a main effect at another locus—say, X—which is not included in the procedure but which is in strong LD with A. Detection of the interaction effect but not a main effect at C, say, may correspond equally to a main effect at X, which—because of LD with A, B, and C—results in a complex interaction A*B*C when X is not included in the model. Indeed, the entire effect of all five loci could potentially be explained by complex patterns of LD with an unknown polymorphism X. (This observation is a feature of the data rather than of the analysis method used; i.e., this would also be true for use of any of the methods previously described, such as the haplotype method, the homozygous parent TDT, the CETDT, or the methods of Cucca et al. [2001*a*] and Zavattari et al. [2001].) We can, however, address this issue by examining the sources of the interaction terms in analysis of a phase-known data set, using method 2, for example. Table 6 shows the results of adding two-way interaction terms to the main effects model A+B+D (with interaction between A and B denoted "A.B," etc.). We find a significant interaction term in either A.D or B.D ($P = .0003$ and $P = .0006$, respectively); once one of these interactions has been included, the other is not significant. The interaction could be independent of phase (an epistatic interaction) or could be restricted to loci on the same chromosome (a haplotype effect). If there is a significant haplotype effect, it would suggest that the interaction effect is in fact partly due to LD with another locus—say, X—not included in the current model. We can test this, in a phase-known data set, by including in the regression procedure an indicator variable, δ_{ij}, that indicates when an individual who is doubly heterozygous, at loci *i* and *j*, has the associated alleles at the two loci occurring on the same parental chromosome—that is, in coupling. This provides a 1-df test of whether there are, in fact, haplotype effects. From table 6, we find that the A.D and B.D interactions do, in fact, each incorporate significant haplotype effects ($P = .002$ and $P = 1 \times 10^{-5}$, respectively). It therefore seems likely that there is another functional polymorphism in the region, apart from A (*DRB1*), D (*DOB*) and B (*DQB1*). This additional polymorphism is unlikely to correspond to C (*DMB*) or E (*TNFc*), since these loci did not have significant main effects in the regression procedure. Indeed, even when the main effects of these loci are included, the terms δ_{AD} and δ_{BD} remain significant (table 6; $P = .003$ and $P = .0002$, respectively), suggesting that there is another functional polymorphism, X, that is not accounted for by any of A, B, C, D, or E. If this polymorphism X were included in the model, it could potentially explain the observed pattern of significance at any or all of A, B, C, D, and E.

Discussion

We have presented a procedure for evaluating the relative contribution to disease of variants within a small genetic region showing strong intermarker LD and for

Table 6

Likelihood-Ratio Tests of Main Effects plus Two-Way Interactions in a Stepwise-Regression Procedure

		METHOD 2		
NULL MODEL	ALTERNATIVE MODEL	χ^2	df	*P*
A+B+D	A+B+D+A.B	5.13	2	.08
A+B+D	A+B+D+A.D	18.61	3	.0003
A+B+D	A+B+D+B.D	17.30	3	.0006
A+B+D+A.D	A+B+D+A.D+B.D	3.12	2	.21
A+B+D+A.D	A+B+D+A.D+ δ_{AD}	9.47	1	.002
A+B+D+B.D	A+B+D+B.D+A.D	4.43	2	.11
A+B+D+B.D	A+B+D+B.D+ δ_{BD}	18.86	1	1×10^{-5}
A+B+C+D+E	A+B+C+D+E+A.B	…	0	…
A+B+C+D+E	A+B+C+D+E+A.C	8.99	4	.06
A+B+C+D+E	A+B+C+D+E+A.D	18.99	2	.0001
A+B+C+D+E	A+B+C+D+E+A.E	2.87	4	.58
A+B+C+D+E	A+B+C+D+E+B.C	7.27	4	.12
A+B+C+D+E	A+B+C+D+E+B.D	17.98	2	.0001
A+B+C+D+E	A+B+C+D+E+B.E	4.31	4	.37
A+B+C+D+E	A+B+C+D+E+C.D	7.27	4	.12
A+B+C+D+E	A+B+C+D+E+C.E	4.96	4	.29
A+B+C+D+E	A+B+C+D+E+D.E	2.00	3	.57
A+B+C+D+E+A.D	A+B+C+D+E+A.D+B.D	2.94	2	.23
A+B+C+D+E+A.D	A+B+C+D+E+A.D+ δ_{AD}	8.60	1	.003
A+B+C+D+E+B.D	A+B+C+D+E+B.D+A.D	3.94	3	.27
A+B+C+D+E+B.D	A+B+C+D+E+B.D+ δ_{BD}	18.26	1	.0002

NOTE.—A = *DRB1*, B = *DQB1*, C = *DMB*, D = *DOB*, and E = *TNFc*. δ_{ij} refers to the variable measuring haplotype effects (see text).

determining which of these variants are likely to be of primary etiological importance. Although the statistical methodology described is not new, very rarely have such stepwise regression procedures been applied in the context of analysis of multilocus haplotypes or genotypes in the human genetics literature. This is surprising, since these methods offer considerable advantages over tests that simply examine association between genotype and disease at each locus in turn, while stratifying by effects at other loci. The regression approach allows the effect at a locus to be examined, conditional on whatever alleles are present at the other loci, without assuming particular values for these alleles. By fitting different models that include main effects and/or statistical interactions, a high degree of flexibility can be achieved for testing a wide variety of null hypotheses. This model framework includes as a subset tests that have been previously proposed in the literature; for example, when method 1 is used and multiplicative effects are assumed, tests of full nested interaction models (such as A*B*C vs. A*B) are essentially equivalent to the CETDT proposed by Koeleman et al. (2000). Such tests typically have large numbers of degrees of freedom because of the large number of estimated parameters, and use of main-effects models instead can offer advantages in terms of the fewer degrees of freedom required for each individual test. The regression approach also allows modeling of full genotype effects rather than simply allelic effects (which rely on the multiplicative assumption, to consider individual parental contributions separately). Consideration of individual parental haplotypes does have the advantage of allowing inclusion of families where one parent is typed but the other is untyped at some or all markers. However, care is needed to avoid bias when including such families (Knapp 1999).

Construction of haplotypes in family data is not a trivial problem, either conceptually or computationally. Although software tools for studying haplotypic associations are now emerging (Clayton 1999; Dudbridge et al. 2000), there are still numerous issues concerning the number of loci to be used for haplotype determination, whether recoding of alleles may be required, and the unbiased treatment of missing data or ambiguous haplotypes (Knapp 1999; Dudbridge et al. 2000). Methods that involve determination of haplotypes (such as methods 1 and 2 described here) have the advantage that they can be used to fit models for the full genotype and haplotype interaction effects of loci. This can be of interest for characterization of the genetic effects or the pattern of LD in a region. However, for detection of loci of primary functional importance (i.e., regardless of other effects in the region), genotypes rather than haplotypes are likely to be most relevant. We therefore propose a method (method 3) that allows analysis of genotype effects only, without determination of the haplotypes present in the data set. This method has the

advantage of making full use of the available families, requiring only that parents and affected offspring be typed at all loci of interest, not that parental phase be known. Method 3 is computationally trivial compared with methods 1 and 2, requiring consideration of each locus individually, rather than in haplotype combination, for the construction of cases and pseudocontrols. However, method 3 does not make use of the information from three (as opposed to one) pseudocontrols, when these are available. We therefore also propose an additional method (method 4), which consists of using method 2 for construction of pseudocontrols in families where phase is inferable and method 3 in families where phase is not inferable.

In genotype analysis of real data for type 1 diabetes, methods 2-4 performed similarly in a data set with polymorphic markers in which parental phase was mostly inferable. Our analysis of the *HLA-DRB1*, *HLA-DQB1*, *DMB*, *DOB*, and *TNFc* loci in type 1 diabetic families confirmed the results of Zavattari et al. (2001), suggesting that each of these loci is important in causing disease, even when simultaneously accounting for effects at the other loci. However, only loci *DRB1*, *DOB*, and possibly *DQB1* appeared to have primary effects in their own right; in these data, *DMB* and *TNFc* appeared to act either through modifying/interaction effects on the other loci or through LD with an etiological locus (or loci) not included in the current analysis. The low relative contribution of *DQB1* in this data set illustrates the impact of population-specific factors, such as allele and haplotype frequencies, on the power to discriminate between variants in tight LD, and it highlights the importance of comparing the results from these types of analysis in distantly related populations.

Needless to say, having untyped loci in the region of LD is a handicap in attempts to identify the etiological loci. Indeed, on the basis of our mathematical modeling alone, the *DRB1*, *DQB1*, and *DOB* effects could in theory be explained by a single locus (or loci) in LD with them that was not included in the analysis. However, on the basis of a substantial body of genetic, functional, and structural data (Cucca et al. 2001*b*), it is well established that *DRB1* and *DQB1* are primary etiological loci in type 1 diabetes. This leaves the *DOB* effect, which is not explained by *DRB1* and *DQB1* or by *DMB* or *TNFc*. The *DOB*-associated effect and the effects associated with *DMB* and *TNFc* could be explained by one (or more) locus not yet typed. Since *DOB* and *DMB*, which are in the class II region, are centromeric of a hot spot of recombination between *DOB* and *DQB1* (Zavattari et al. 2000), and given mapping data from other studies (Ota et al. 1999; Nejentsev et al. 2000), our results for this Sardinian data set would be consistent with an untyped etiological locus in the class II region and one other untyped locus, perhaps in the class III region near *TNFc*. The class II locus could be the third classical antigen-presentation gene, *HLA-DPB1*, which is also associated with type 1 diabetes in a primary way (Cucca et al. 2001*a*). However, this is unlikely, since the associated *DPB1*0402* allele is only present in 3.6% of Sardinian *DR3* haplotypes and, therefore, cannot account for the non-*DR-DQ* effect observed here (Zavattari et al. 2001). The putative class III locus could be the same type 1 diabetes locus mapped, in Finnish subjects, to 240 kb of the class III region near *HLA-B* at the telomeric end of the class II region (Nejentsev et al. 2000). Furthermore, it could correspond to an autoimmune disease locus mapped to 46 kb of this same 240-kb segment of the class III region (Ota et al. 1999; Nejentsev et al. 2000). The highly polymorphic immune response molecules, MICA and MICB, which are encoded in the segment, with MICA in the 46-kb region, are prime candidates for a class III disease locus. Full typing of these and other genes in this specific region is a next step in the identification of non-*DR/DQ* loci in type 1 diabetes and in other *HLA*-associated diseases (Gambelunghe et al. 1999).

Acknowledgments

We thank Frank Dudbridge and John Todd for useful discussions and Francesco Cucca for providing the data from the Zavattari et al. (2001) paper. Support for the authors is provided by two Wellcome Trust Research Fellowships jointly funded by the Wellcome Trust and the Juvenile Diabetes Research Foundation.

Electronic-Database Information

The URL for software used in this study is as follows:

Authors' Web site, http://www-gene.cimr.cam.ac.uk/clayton/software/stata/ (for software for fitting the models described here in the statistical package Stata)

References

Beck S, Trowsdale J (1999) Sequence organisation of the class II region of the human MHC. Immunol Rev 167:201–210

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

Cucca F, Dudbridge F, Loddo M, Mulargia AP, Lampis R, Angius E, De Virgiliis S, Koeleman BP, Bain SC, Barnett AH, Gilchrist F, Cordell H, Welsh K, Todd JA (2001*a*) The HLA-DPB1–associated component of the IDDM1 and its relationship to the major loci HLA-DQB1, -DQA1, and -DRB1. Diabetes 50:1200–1205

Cucca F, Lampis R, Congia M, Angius E, Nutland S, Bain SC, Barnett AH, Todd JA (2001*b*) A correlation between the relative predisposition of MHC class II alleles to type 1

diabetes and the structure of their proteins. Hum Mol Genet 10:2025–2037

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). J R Stat Soc B 39:1–38

Dudbridge F, Koeleman BPC, Todd JA, Clayton DG (2000) Unbiased application of the transmission/disequilibrium test to multilocus haplotypes. Am J Hum Genet 66:2009–2012

Falk CT, Rubenstein P (1987) Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. Ann Hum Genet 51:227–233

Gambelunghe G, Falorni A, Ghaderi M, Laureti S, Tortoioli C, Santeusanio F, Brunetti P, Sanjeevi CB (1999) Microsatellite polymorphism of the MHC class I chain-related (MIC-A and MIC-B) genes marks the risk for autoimmune Addison's disease. J Clin Endocrinol Metab 84:3701–3707

Hirji KF, Mehta CR, Patel NR (1987) Computing distributions for exact logistic regression. J Am Stat Assoc 82:1110–1117

——— (1988) Exact inference for case-control studies. Biometrics 44:804–814

Huber P (1967) The behaviour of maximum likelihood estimates under non-standard conditions. In: Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics and Probability. Vol 1. University of California Press, Berkeley, pp 221–233

Knapp M (1999) The transmission/disequilibrium test and parental-genotype reconstruction: the reconstruction-combined transmission/disequilibrium test. Am J Hum Genet 64: 861–870

Koeleman BP, Dudbridge F, Cordell HJ, Todd JA (2000) Adaptation of the extended transmission/disequilibrium test to distinguish disease associations of multiple loci: the Conditional Extended Transmission/Disequilibrium Test. Ann Hum Genet 64:207–213

Li H (2001) A permutation procedure for the haplotype method for identification of disease-predisposing variants. Ann Hum Genet 65:189–196

Lie BA, Todd JA, Pociot F, Nerup J, Akelsen HE, Joner G, Dahl-Jorgensen K, Ronningen KS, Thorsby E, Undlien DE (1999) The predisposition to type 1 diabetes linked to the human leukocyte antigen complex includes at least one non-class II gene. Am J Hum Genet 64:793–800

Martin ER, Kaplan NL, Weir BS (1997) Tests for linkage and association in nuclear families. Am J Hum Genet 61:439–448

McCullagh P, Nelder JA (1989) Generalized linear models, 2d ed. Chapman & Hall, London

Mignot E, Lin L, Rogers W, Honda Y, Qiu X, Lin X, Okun M, Hohjoh H, Miki T, Hsu S, Leffell M, Grumet F, Fernandez-Vina M, Honda M, Risch N (2001) Complex HLA-DR and -DQ interactions confer risk of narcolepsy-cataplexy in three ethnic groups. Am J Hum Genet 68:686–699

Nejentsev S, Gombos Z, Laine AP, Veijola R, Knip M, Simell O, Vaarala O, Akerblom HK, Ilonen J (2000) Non-class II HLA gene associated with type 1 diabetes maps to the 240-kb region near HLA-B. Diabetes 49:2217–2221

Nelder JA, Wedderburn RWM (1972) Generalized linear models. J R Stat Soc A 135:370–384

Ota M, Mizuki N, Katsuyama Y, Tamiya G, Shiina T, Oka A, Ando H, Kimura M, Goto K, Ohno S, Inoko H (1999) The critical region for Behçet disease in the human major histocompatibility complex is reduced to a 46-kb segment centromeric of HLA-B, by association analysis using refined microsatellite mapping. Am J Hum Genet 64:1406–1410

Robinson WP, Barbosa J, Rich SS, Thomson G (1993) Homozygous parent affected sib pair method for detecting disease predisposing variants: application to insulin dependent diabetes mellitus. Genet Epidemiol 10:273–288

Schafer JL (1997) Analysis of incomplete multivariate data. Chapman & Hall, London

Schaid DJ (1996) General score tests for associations of genetic markers with disease using cases and their parents. Genet Epidemiol 13:423–449

Self SG, Longton G, Kopecky KJ, Liang K-Y (1991) On estimating HLA/disease association with application to a study of aplastic anemia. Biometrics 47:53–61

Thomas D, Pitkaniemi J, Langholz B, Tuomilehto-Wolf E, Tuomilehto J (1995) Variation in HLA-associated risks of childhood insulin-dependent diabetes in the Finnish population: II. Haplotype effects. DiMe Study Group. Childhood Diabetes in Finland. Genet Epidemiol 12:455–466

Thomson G (1995) Mapping disease genes: family-based association studies. Am J Hum Genet 57:487–498

Thomson G, Robinson WP, Kuhner MK, Joe S, MacDonald MJ, Gottschall JL, Barbosa J, Rich SS, Bertrams J, Baur MP, Partanen J, Tait BD, Schober E, Mayr WR, Ludvigsson J, Lindblom B, Farid NR, Thompson C, Deschamps I (1988) Genetic heterogeneity, modes of inheritance, and risk estimates for a joint study of Caucasians with insulin-dependent diabetes mellitus. Am J Hum Genet 43:799–816

Valdes AM, Thomson G (1997) Detecting disease-predisposing variants: the haplotype method. Am J Hum Genet 60:703–716

White H (1982) Maximum likelihood estimation of misspecified models. Econometrica 50:1–25

Zavattari P, Deidda E, Whalen M, Lampis R, Mulargia A, Loddo M, Eaves I, Mastio G, Todd JA, Cucca F (2000) Major factors influencing linkage disequilibrium by analysis of different chromosome regions in distinct populations: demography, chromosome recombination frequency and selection. Hum Mol Genet 9:2947–2957

Zavattari P, Lampis R, Motzo C, Loddo M, Mulargia A, Whalen M, Maioli M, Angius E, Todd JA, Cucca F (2001) Conditional linkage disequilibrium analysis of a complex disease superlocus, IDDM1 in the HLA region, reveals the presence of independent modifying gene effects influencing the type 1 diabetes risk encoded by the major HLA-DQB1, -DRB1 disease loci. Hum Mol Genet 10:881–889