doi: 10.1111/j.1469-1809.2005.00218.x

Case-parent Triads: Estimating Single- and Double-dose Effects of Fetal and Maternal Disease Gene Haplotypes

H. K. GJESSING[1,2*] and R. T. LIE[2,1]

[1] *Division of Epidemiology, Norwegian Institute of Public Health, Norway*

[2] *Section for Epidermiology and Medical Statistics, Department of Public Health and Primary Health Care, University of Bergen, Norway*

Summary

Case-parent triad data are considered a robust basis for studying association between variants of a gene and a disease. Methods evaluating statistical significance of association, like the TDT-test and its extensions, are frequently used. When there are prior hypotheses of a causal effect of the gene under study, however, methods measuring penetrance of alleles or haplotypes as relative risks will be more informative. Log-linear models have been proposed as a flexible tool for such relative risk estimation. We demonstrate an extension of the log-linear model to a natural framework for also estimating effects of multiple alleles or haplotypes, incorporating both single- and double-dose effects. The model also incorporates effects of single- and double-dose maternal haplotypes on a fetus during pregnancy. Unknown phase of haplotypes as well as missing parents are accounted for by the EM algorithm. A number of numerical improvements to maximum likelihood estimation are also implemented to facilitate a larger number of haplotypes. Software for these analyses, HAPLIN, is publicly available through our web site. As an illustration we have re-analyzed data on the MSX1 homeobox-gene on chromosome 4 to show how haplotypes may influence the risk of oral clefts.

Introduction

Since Falk & Rubinstein (1987) and Self *et al.* (1991) proposed that genotypes of parents of cases could be used to study association between disease and allelic variants, and Spielman *et al.* (1993) introduced the transmission disequilibrium test (TDT), the case-parent triad design has become an increasingly important approach for association studies. The standard case-parent triad design is based on selecting case children from a population, and then genotyping both children and their parents. Since the case-parent triad design has strengths and weaknesses that are different from those of the case-control design (Weinberg & Umbach, 2000), case-triad studies are also an important method of verification of association detected by case-control studies (NatureGenetics, 1999). Assessment of replication of association between case-parent and case-control studies requires that a measure of association (e.g. relative risk) is available from both designs (Mitchell 2000). A substantial part of the literature on case-parent triad data is, however, dedicated to calculating p-values using the TDT-test or related tests (Clayton, 1999), (Laird, 2000), (Zhao, 2000a,b), (Dudbridge, 2003), (Horvath, 2004). Attractive methods based on log-linear models are available for estimation of relative risk for diallelic markers (Weinberg, 1998), (Wilcox, 1998), (Weinberg, 1999b), (Umbach, 2000b). The basis of the log-linear model application is to list all possible triad genotypes, and applying a log-linear model for the frequencies of the different triad types conditional on the child being a case. The log-linear model has a number of convenient features. First, it produces relative risk estimates for a single or double dose of a deleterious allele, rather than just a hypothesis test. Second, it deals with incomplete triad data in a relatively straightforward manner (Weinberg, 1999a). Third, it can incorporate other types of effect estimates

*Address for correspondence: Dr. Håkon K. Gjessing, Norwegian Institute of Public Health, P.O. Box 4404 Nydalen, N-0403 Oslo, Norway. Phone: +47 23408241, Fax: +47 23408260. E-mail: hakon.gjessing@fhi.no

than just the direct effect of the child's alleles. Gene-environment interactions as well as effects of maternal genes can be incorporated (Wilcox, 1998), (Umbach & Weinberg, 2000b). The basic model is parametrized either using a Hardy-Weinberg equilibrium assumption for the parental generation, or with completely unrestricted mating type frequencies.

The possibility of extending the log-linear model to a situation with multiple alleles at a locus is of particular relevance. Perhaps the most immediate application of this is when studying haplotypes, which are inherently polymorphic. The ongoing effort of identifying single nucleotide polymorphisms (SNPs) in the human genome provides about 15 SNPs for an average-length functional locus (International SNP map working group, 2001), (International human genomesequencing consortium, 2001). Although $2^{15} = 32,768$ different haplotypes are in principle possible at such a locus, a fairly small number is actually seen in practice, revealing strong linkage disequilibrium. At present, there are a number of difficulties with applying the log-linear model to situations with a potentially vast number of haplotypes. First, for SNP data with unknown phase the haplotypes must be reconstructed from parental data whenever possible, and haplotype frequencies must be predicted from the model for the remaining triads. Second, a full model taking into account all possible genotype effects will contain too many parameters to be practical. Several of the many haplotypes could conceivably confer an elevated risk, and there could be complex interaction patterns between two haplotypes at the same locus. Third, a standard application of log-linear software quickly becomes infeasible, since the number of possible triads becomes unmanageable even with a moderate number of SNPs. For instance, 15 SNPs will produce 1.1×10^{18} possible triads, making a full enumeration impossible.

Recently, a related but alternative approach to log-linear modelling was extended to estimation of relative risks associated with haplotypes (Cordell, 2004a, 2004b). The approach applies the principle of conditioning on parental mating type (Self *et al.* 1991), (Terwilliger & Ott, 1992), (Clayton, 1999), (Sham & Curtis, 1995), using the case alleles and "pseudo-controls" constructed from the non-transmitted alleles (Schaid, 1996, Khoury, 1994). An advantage of the conditioning approach is that it does not depend on assumptions about population structure, like Hardy-Weinberg equilibrium (HWE). Furthermore, the analyses can be done using conditional logistic regression software. On the other hand, it is well known that the conditioning decreases power in situations where HWE can be assumed (Knapp *et al.* 1995). In addition, a direct application of the pseudocontrol approach requires discarding triads where transmission is ambiguous, i.e. when parent-of-origin is unobservable. This problem extends similarly to SNPs where phase is unknown, leading to a greater loss of data when many SNPs are involved.

In the present work we extend the log-linear model to a locus with multiple alleles or haplotypes with unknown phase. We suggest a parametrization that allows for reasonable flexibility without attempting to estimate too many parameters. The model estimates both fetal single- and double-dose haplotype relative risks for all haplotypes, using the remaining haplotypes (or alternatively a single haplotype) as reference. It thus also allows an assessment of whether there is a dose-response pattern, or a recessive or dominant effect. In a similar fashion, effects of maternal haplotypes are estimated, and parent-of-origin effects are considered. The model is based on a full maximum likelihood approach and standard likelihood ratio tests are available to compare nested models. We also suggest a number of computational simplifications that substantially reduce the numerical problems. An example demonstrates how the model can be applied to estimate haplotype relative risks of cleft lip or palate for the MSX1 homeobox gene on chromosome 4, where no deviation from HWE was seen. Our software, HAPLIN, is developed to estimate these models. It includes the parametrizations described in this article for fetal and maternal haplotypes. The EM (expectation maximization) algorithm is supplemented with jackknife resampling to estimate standard errors and confidence intervals. HAPLIN is available from our web site.

Log-linear Models for a Multi-Allelic Locus in Hardy-Weinberg Equilibrium

Notation

Consider a single locus with K alleles $A_1, A_2, \ldots, A_K$ and with population allele frequencies $p_1, p_2, \ldots, p_K$.

Let M, F and C denote the genotypes for the mother, the father and the child, respectively, and let (M, F) be the corresponding mating types and (M, F, C) the triad type. For instance, if $M = A_1A_2$, $F = A_2A_3$ and $C = A_2A_3$ the mating type is written $(M, F) = (A_1A_2, A_2A_3) = A_1A_2 \times A_2A_3$ and the triad type is $(M, F, C) = (A_1A_2, A_2A_3, A_2A_3)$. We will follow a strict ordering in that the first genotype belongs to the mother and the second to the father. Also, we assume that the second allele form the mother and the second allele from the father are inherited by the child. Thus, in the triad type (A_1A_2, A_2A_3, A_2A_3) allele A_2 comes from the mother and A_3 from the father. With this notational convention, specification of the mating type (A_1A_2, A_2A_3) is used to describe the full triad.

Furthermore, let n_{ijkl} be the frequency of the triad type (A_iA_j, A_kA_l) in the observed data, $1 \leq i, j, k, l \leq K$. Note that when all three individuals are heterozygous for the same two alleles, i.e. $i = l \neq j = k$, only the sum $n_{ijji} + n_{jiij}$ can be observed directly from the data, not the two separate frequencies. However, as long as no parent-of-origin effects are included, this ambiguity is irrelevant for the likelihood, and the sum frequency can be distributed arbitrarily among the two groups.

We will use for instance $n_{i\cdots} = \sum_{jkl} n_{ijkl}$ to denote summation over indices, and $n = n_{\cdots\cdot}$ for the total sample size.

Sampling Model

Let D denote the event that the child has the disease. A triad is sampled through a case child. The information n_{ijkl} in the observed data thus relates to the triad probabilities $P(M, F, C|D)$, conditional on disease in the child. Under this conditioning, we assume a Poisson probability model for the triad type frequencies n_{ijkl}, with expected cell values proportional to $P(M, F, C|D)$. By a standard Bayes argument we write

$$P(M, F, C|D) = P(D|M, F, C)P(M, F, C)/P(D). \tag{1}$$

The disease prevalence $P(D)$ enters the model only as a normalizing constant, unidentifiable due to the sampling scheme. The triad population frequencies $P(M, F, C)$ are typically considered "nuisance" parameters, whereas the disease penetrance, $P(D|M, F, C)$, where the effect of genotype on risk is modelled, is an essential part of the model. We will consider the separate parts below.

Triad Frequencies

The triad population frequencies can be decomposed as

$$P(M, F, C) = P(C|M, F)P(M, F).$$

The transmission probability part $P(C|M, F)$ is trivial when assuming Mendelian transmission. The mating type frequencies $P(M, F)$ are population quantities, depending on population structure, mating pattern etc. We assume that the population is in Hardy-Weinberg equilibrium at the locus, thus assuming, among other things, that there is random mating and no population stratification. Based on these assumptions, together with Mendelian transmission, we can express the triad frequencies simply as

$$P(M, F, C) = P(A_iA_j, A_kA_l) = p_i p_j p_k p_l. \tag{2}$$

Note that this is an assumption about the unselected population. We do not assume the group of case children to be in HWE. In fact, if a deleterious effect of the genes under study exists, the case group will only be in HWE if the genetic effect is multiplicative (Lee, 2003).

Disease Penetrance

The most important modelling decision lies in how to represent $P(D|M, F, C)$, i.e. the risk of a child exhibiting the disease, conditional on the triad genotype. The simplest versions appear when we assume information about the mating type (M, F) is irrelevant when the child's genotype C is known. We may then write $P(D|M, F, C) = P(D|C)$, and focus on how the genotype of the child directly influences the risk of disease. However, particularly the maternal genotype M may be thought to influence the development of the child as a fetus, and is thus often considered in models in perinatal epidemiology (Wilcox *et al.* 1998), (Cordell *et al.* 2004b). Furthermore, parent-of-origin effects are also identifiable from the full triad genotype but not from the genotype of the child alone. Yet another situmaternal and fetal alleles (Sinsheimer *et al.* 2003). Finally, if the gene under study is not the disease gene itself, nor in very close linkage to it, the genotype of the child alone may not contain all information relevant to the allele distribution at the disease locus (Weinberg, 1999b, Cordell, 2004a).

Below, we will consider different possible models for how the child- and parental genotypes influence the probability of disease.

Single- and Double-dose Effects

We start by looking at different parametrizations when we assume the parental genotypes can be disregarded, i.e. when $P(D|M, F, C) = P(D|C) = P(D|A_j A_l)$. Let us for the moment consider only the diallelic situation, with alleles A

model; the relative risk obtained when comparing A_jA_x individuals with A_kA_x individuals is R_j/R_k for all A_x, *provided that* $x \neq j, k$. That is, the comparison is made only among heterozygotes; homozygous individuals are given separate parameters. We see that the A_1A_1 homozygotes can no longer be thought of as a reference group since they have the risk $\eta\tilde{R}_1$, which may differ from the baseline level η. Even when $R_1 = 1$ the parameter $\tilde{R}_1$ should still be estimated and will usually differ from 1. The baseline level η corresponds to the risk of an A_1 homozygote *only* if the multiplicative model holds. However, A_1 still retains its status as reference *allele*.

The reason we prefer the reference allele approach to using A_1 homozygotes as a reference group is that the single dose effects $R_2, \ldots, R_K$ are frequently more precisely estimated than the double-dose effects $\tilde{R}_1, \ldots, \tilde{R}_K$; even for the reference allele homozygotes may be relatively scarce and provide an unstable basis for a reference level. It should be noted that for the diallelic situation the two approaches coincide since we cannot estimate $\tilde{R}_1$ separately and must set it equal to one.

The double-dose estimates will provide an impression of the effect of allele dose on risk. For the effect of A_2 one can compare the estimates $\tilde{R}_1$, R_2 and $\tilde{R}_2$. If $\tilde{R}_1 \approx R_2$ there is a recessive effect of A_2, if $R_2 \approx \tilde{R}_2$ there is a dominant effect and if $R_2/\tilde{R}_1 \approx \tilde{R}_2/R_2$ there is a dose-response relationship between A_1 and A_2. For this reason, it is instructive to create a plot of $\tilde{R}_1$ together with R_j and $\tilde{R}_j$ for the remaining alleles, $j = 2, \ldots, K$, thus enabling an easy visual inspection of possible penetrance patterns. It should be kept in mind, however, that this depends on A_1 being the reference. We cannot directly evaluate the relationship between, for instance, A_2 and A_3 in the same manner without changing the reference allele.

Reciprocal Reference

In some situations there may not be any natural candidates as reference alleles; perhaps there is no preconceived idea about which allele is the wild type and which allele is deleterious. One might face a situation where, say, alleles A_1, A_2 and A_3 are neutral whereas A_4 is deleterious, and none of the alleles A_1, A_2 and A_3 are particularly more frequent than the others. It would then seem artificial to choose A_1 as reference since this would result in a comparison only between A_4 and A_1, whereas one would prefer to compare A_4 with A_1, A_2 and A_3 jointly; this would provide a wider (and thus more stable) basis for the reference, thus improving power for the comparison. Additionally, as seen in the previous subsection, a discussion of penetrance patterns is always relative to a reference category. It may be more relevant to ask whether A_4 on the average is dominant or recessive relative to the collective of wild type alleles, rather than to an arbitrarily selected reference allele. This can be achieved by using a "reciprocal" reference, meaning that for each allele all the remaining alleles are used as a joint reference.

Let P_i be the probability of disease for a heterozygous individual picked at random among individuals with *exactly one* A_i-allele, and let P_{i-} be the same probability for an individual picked at random from all heterozygotes *not* having any A_i-alleles. When using the reciprocal reference we estimate parameters from (3) as before, but then compute new single- and double-dose estimates as $F_i = P_i/P_{i-}$ and $\tilde{F}_i = \eta\tilde{R}_i/P_{i-}$. Both P_i and P_{i-} can be computed from the parameters in (3), up to the baseline probability η. The unknown η cancels out in the expressions for F_i and $\tilde{F}_i$. The interpretation of F_i is then the increase (or decrease) in risk seen when picking a random heterozygote without the A_i-allele and replacing one of the alleles with an A_i-allele. For allele A_i the reference is thus all heterozygotes not carrying the A_i-allele. Similarly, $\tilde{F}_i$ is the change in risk seen when replacing *both* alleles with A_i-alleles. Note that with the reciprocal reference the reference can be interpreted both as a collection of alleles (those different from A_i) and as a group of individuals (the heterozygotes not carrying A_i). The only disadvantage is that the reference category depends on which allele is under study. For this parametrization, one may present a table or plot of both F_i and $\tilde{F}_i$ for all alleles. A recessive effect is seen when $F_i = 1$ and $\tilde{F}_i$ is significantly different from 1. Similarly, $F_i = \tilde{F}_i$ indicates a dominant effect, and $\tilde{F}_i = F_i^2$ a dose-response. It should be kept in mind, however, that concepts such as dominance and recessiveness are always relative to a reference, which in this case is a composite group of alleles. Thus, it should be seen as an average effect rather than something absolute.

The Log-linear Model

For notational convenience we will, in the following, rewrite the penetrance model (3) as $P(D|A_jA_l) = \eta R_j R_l R^\star_{jl}$ where $R^\star_{jl} = R^\star_j$ when $j = l$ and $R^\star_{jl} = 1$ when $j \neq l$. The two model formulations are equivalent. Let $\xi_{ijkl} = E[n_{ijkl}]$ be the expected (conditional) frequency of triads of type (A_iA_j, A_kA_l). Entering (2) and (3) into (1), the expected triad frequencies are

$$\begin{aligned}\xi_{ijkl} &= \xi \cdot P((A_i A_j, A_k A_l)|D) \\ &= \xi \cdot p_i p_j p_k p_l \cdot R_j R_l \cdot R^\star_{jl} \qquad (4)\end{aligned}$$

where ξ is a normalizing constant. Due to the multiplicative structure we can write

$$\log(\xi_{ijkl}) = \mathbb{X}_1 \boldsymbol{\beta}_1 + \mathbb{X}_2 \boldsymbol{\beta}_2 + \mathbb{X}_3 \boldsymbol{\beta}_3 \qquad (5)$$

where $\boldsymbol{\beta}_1 = (\log(p_1), \ldots, \log(p_K))^T$, $\boldsymbol{\beta}_2 = (\log(R_2), \ldots, \log(R_K))^T$, $\boldsymbol{\beta}_3 = (\log(R^\star_1), \ldots, \log(R^\star_K))^T$, and $\mathbb{X}_1, \mathbb{X}_2$ and $\mathbb{X}_3$ are appropriate design matrices of dimensions $K^4 \times K, K^4 \times (K-1)$ and $K^4 \times K$, respectively. Note that ξ is incorporated by first estimating $\boldsymbol{\beta}_1$ freely, disregarding the restriction $\sum_i p_i = 1$, and then recovering the allele frequencies from $p_i = \exp(\beta_{1i}) / \sum_j \exp(\beta_{1j})$, where β_{1i} are the components of $\boldsymbol{\beta}_1$. Thus, assuming a Poisson likelihood in the maximum likelihood estimation (MLE), this is a log-linear model where each triad type contributes the term

$$n_{ijkl} \log \xi_{ijkl} - \xi_{ijkl} \qquad (6)$$

to the log-likelihood (up to an additive constant), where ξ_{ijkl} is given in (4).

For the fully multiplicative model, i.e. when $R^\star_j = 1, j = 1, \ldots, K$, there are explicit solutions to the likelihood equations (see Appendix), and the risk parametrization is parallel to that of the gamete-competition models (Sinsheimer, 2000). When including the double-dose effects, the numerical solutions cease to be completely explicit. Nevertheless, a near-explicit solution can be found, necessitating the estimation of only one parameter, from which all other parameter estimates can be computed. In addition, the EM algorithm provides a simple framework that quite easily extends the explicit solution for the multiplicative model to the situation where double-dose effects are included. Both approaches are discussed in detail in the appendix. For more information about general use of the EM algorithm in genetics see, e.g., (Sorensen & Gianola 2002).

Since the model is based on a full maximum likelihood estimation, standard likelihood ratio tests can be performed to compare nested models. For instance, one might test whether $R^\star_j = 1$ for all j, so that the model could be reduced to the fully multiplicative one. Similarly, a joint test for effects of maternal alleles (see below) could be performed, or an overall test for the effect of all alleles at the locus. All tests are based on the likelihood ratio, with a chi-squared distribution as the asymptotic null distribution and using the added number of parameters in the largest model as the degrees of freedom. These tests are also valid when the EM algorithm is used to maximize the likelihood, provided the likelihood is computed in the original model with unobserved information. In addition, Wald-based p-values for effects of individual alleles can be provided, avoiding having to compute the maximum likelihood estimates for all submodels.

Effects of Maternal Alleles

For a model including the possible effect of the maternal alleles, it is natural to use the same parameterization as for the fetal alleles, and assume that the maternal alleles have a multiplicative effect in addition to the fetal alleles. This results in a model

$$P(D|(A_i A_j, A_k A_l)) = \eta \cdot R_j R_l R^\star_{jl} \cdot M_i M_j M^\star_{ij}, \qquad (7)$$

where the parameters $M_1, M_2, \ldots, M_K$ and $M^\star_1, M^\star_2, \ldots, M^\star_K$ have an interpretation similar to their fetal counterparts.

Recall that our model assumes Hardy-Weinberg equilibrium/random mating. Inherent in this assumption is the assumption of "mating symmetry" between the mother and the father, in the sense that the allele population frequencies are the same for males and females. Whereas this assumption is likely to be less critical when studying the effects of *fetal* alleles, it is crucial when estimating the effect of *maternal* alleles during pregnancy. In effect, equation (7) relies on a contrast between the allele frequencies for the mother and those for the father when estimating the effect of maternal alleles. This may be questionable in, for instance, populations where a substantial number of marriages are between males from the local population and female immigrants, or vice versa.

Other Effects

Under a mild assumption about independence of exposure and child genotype conditional on parental mating type (Umbach & Weinberg, 2000b), (Thomas, 2000), interactions between genes and a categorical exposure variable can be included in the model. Essentially, this amounts to fitting separate models for each exposure category, and adds little in terms of computational difficulties.

Under the assumption that the locus under study is functionally related to the disease, parent-of-origin effects can be included (Weinberg, 1999b), (Cordell *et al.* 2004b). In our setting, we can choose to assign different effects to the alleles in the child depending on whether they derive from the mother or the father. This can be accomplished by, for instance, setting

$$P(D|(A_i A_j, A_k A_l)) = \eta \cdot R_j^{(M)} R_l^{(F)} R_{jl}^{\star} \cdot M_i M_j M_{ij}^{\star}, \tag{8}$$

the only difference from (7) being that the single-dose effects R_j, R_l are separated into $R_j^{(M)}$ and $R_l^{(F)}$ depending on whether the allele is derived from the mother or from the father. The fraction $R_j^{(M)}/R_j^{(F)}$ is a measure of how much higher or lower the risk associated with allele A_j is, depending on whether it is transmitted from the mother or the father. We should keep in mind, however, that the model (8) requires knowledge of parent-of-origin, which is not the case for ambiguous triads. For this reason, the parent-of-origin model must be combined with the EM algorithm to estimate the frequency distribution within ambiguous triads, or for haplotypes with unknown phase. Implementation of the EM algorithm is described in more detail below.

Other effects, such as the effects of several unlinked loci and gene-gene interactions, can also be implemented in the log-linear model; we will not go into details here.

In passing, we also note that specific deviations from HWE could be modelled. Since both population stratification and inbreeding typically lead to a deficiency in heterozygotes, one could include a multiplicative parameter allowing for homozygotes to have a higher frequency in the population than expected from HWE. This would lead to a model for the triad frequencies of the form

$$P(M, F, C) = p_i p_j p_{ij}^{\star} \cdot p_k p_l p_{kl}^{\star}$$

where $p_{ii}^{\star} = p_i^{\star}$ has a separate value for each homozygote $A_i A_i$, and $p_{ij}^{\star} = 1$ for the heterozygotes ($i \neq j$). We assume random mating in the last generation, although this is not necessary. Models for deviations from HWE will not be pursued any further in this paper.

Software Implementation

In the diallelic situation it has been shown how the models can be implemented in standard software for log-linear modeling (Weinberg *et al.* 1998). When investigating multiple gene variants, and in particular haplotypes (as described below), there is a need for dedicated software implementations. The models described in this paper are part of our software HAPLIN. It computes both relative risks with confidence intervals and likelihood-ratio and Wald-based p-values for various tests, and presents results both as tables and as figures. HAPLIN allows in particular the effects of maternal alleles to be included. It also provides estimates of allele- and haplotype frequencies with confidence intervals. Several different reference category methods are implemented, reciprocal reference being the default.

Haplotypes and Missing Information

Estimating Haplotype Relative Risk

The application of the log-linear model to multiple alleles described above can now be adapted in a fairly straightforward manner to the situation with multiple closely linked markers within a locus, where phase may be unknown. We will in the following assume that the markers are so strongly linked that recombination hardly ever occurs between them in the transmission from parents to child in our triads.

If phase were known, each haplotype could be treated as a single allele, and estimation could proceed as above. The problem is thus to deduce phase for all markers in all three individuals. Recall that for an ambiguous marker, i.e. a marker with triad genotype (A_iA_j, A_jA_i), the parent of origin cannot be deduced, but for all other types of markers the parent of origin *can* be deduced. When looking at a single triad, if *all* markers were non-ambiguous we could deduce precisely which alleles were transmitted from the mother and which from the father at all markers. All alleles transmitted from the mother must then constitute a haplotype in the child, and similarly those transmitted from the father. We would thus be able to deduce all six haplotypes in the triad. If one or more of the markers happen to be ambiguous, they cannot be linked to the transmitted alleles at the other markers, and we cannot deduce any of the haplotypes in the triad. (The only exception to this rule is when an individual is homozygous at all markers except a single ambiguous marker.)

Clearly, as the number of markers increases, the number of triads with at least one ambiguous marker will increase. Depending on the data this number may soon become substantial. For instance, in one of our datasets for cleft lip/palate, for a gene with only two SNPs, we found that for about 14% of the triads phase could not be deduced directly (data not shown). As a consequence haplotypes need to be reconstructed statistically for those triads that contain at least one ambiguous marker.

β

The most common way of doing the statistical reconstruction is to use the EM algorithm (Cheng *et al.* 2003), which is particularly well suited for log-linear models. The M-step performs the maximization as described above for multiple alleles, as if all haplotypes were known. In the E-step the observed frequencies for ambiguous triads are redistributed according to values predicted from the model. Since it is sufficient to reconstruct parent of origin for the ambiguous markers to enable reconstruction of all haplotypes, a triad with g ambiguous markers will be represented by 2^g possible different haplotype configurations within that triad, each with a predicted frequency. (For the sake of convenience, the exceptional cases mentioned above, where haplotypes can be found even in the presence of ambiguous markers, can be treated as unknown haplotypes in the EM algorithm. This only leads to a negligible reduction in the speed of convergence.)

The well-known drawback of the EM algorithm is that it does not immediately provide standard error estimates for the estimation results. Although standard errors can be computed in each M-step, these do not account for the extra uncertainty resulting from ambiguous haplotypes. Several ways of providing the extra information needed from the EM algorithm to compute correct asymptotic standard errors are described in Sorensen & Gianola (2002). In HAPLIN, the extra uncertainty is accounted for by an option to use jackknifing of standard errors (Efron & Tibshirani, 1993). This is reasonably efficient since the data are in a tabulated format. The jackknifing requires the removal of each triad, one at a time. However, for triad types that exist in multiple copies removal occurs only once, and the resulting contribution is weighted according to the frequency of that triad type. Thus, the number of jackknife replications needed is usually substantially lower than the sample size. To be explicit, let $\boldsymbol{\beta}$ denote the combined vector of (log-transformed) parameters $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ from (5), and let $\hat{\beta}_{ijkl}$ be the estimate obtained when removing one triad from the n_{ijkl} triads of type (A_iA_j, A_kA_l) (when $n_{ijkl} \geq 1$). The jackknife estimate of the standard error is then

$$SE_{jack}(\hat{\beta}) = \Big[n - 1$$

Missing Parental Genotypes

As shown in Weinberg (1999a), missing genotypic information on parents can easily be incorporated, again using the EM algorithm. In fact, since the EM algorithm is already used to identify haplotypes, no extra steps are needed to account for the missing parental information. In the E-step, the expectations can be computed with both phase ambiguity and unknown parental information at the same time. In applications, the possibility of incorporating incomplete triads can be crucial since parental information, particularly for the father, may frequently be missing, sometimes leading to a serious loss of power if the entire triad must be excluded. A requirement for dealing with missing genotype data in this manner is that we can assume missingness is independent of genotype. For most applications this appears reasonable.

Example and Interpretation of Parameters

To illustrate use of the log-linear models in a situation with two markers, we re-analyzed data from a series of 261 case-parent triads of cleft lip or palate with confirmed fathers from Norway (Jugessur *et al.* 2003). The triads were assayed for two markers in the MSX1 homeoboxgene on chromosome 4. Marker 1, the MSX1-CA, was a CA-repeat with four different variants. Marker 2, MSX1-1.3, was a diallelic SNP. Details of the genetic assays may be found elsewhere (Lidral *et al.* 1997, 1998). Since a previous analysis of MSX1 variants indicated an association for both types of clefts (Lidral *et al.* 1998), our cases constitute a mixed set of both cleft lip and cleft palate. In this material, genotyping was done only on triads with genetic material from all family members, and yielded complete results for all triads. If, as is usually the case, genotyping had failed for some markers in some individuals, the incomplete triads could still have been used as described above. HAPLIN was used for all computations.

With four alleles at one marker (denoted 1,2,3,4) and two at the other (1,2) there are eight possible haplotypes, denoted by 1-1, 1-2, 2-1, . . . ,4-1. Haplotypes 1-1, 2-1 and 3-1 have a total frequency of less than 1% and were excluded from the analysis. The haplotypes 4-1, 1-2, and 3-2 rarely occur as homozygotes (frequencies 0, 4 and 1, respectively). We thus omitted parameters estimating double-dose effects of 4-1 and 3-2. We estimated effects for both fetal and maternal haplotypes simultaneously.

The relative risk estimates for single- and double-dose of both fetal and maternal haplotypes are presented in Table 1. All estimates are supplied with 95% confidence intervals. As an alternative presentation, HAPLIN creates separate plots for the fetal and maternal effects. The plot for the fetal effects is shown in Figure 1. All results are presented with reciprocal reference.

This means that the relative risk (single-dose) estimate corresponding, for instance, to haplotype 4-2, is interpreted as the risk of disease for a heterozygous individual carrying one copy of the haplotype, relative to heterozygotes not having the haplotype. It corresponds to the F_i-parameters defined earlier. It is seen that, perhaps somewhat surprisingly, fetal haplotype 4-2 carries a risk above the other haplotypes, even though this is by far the most frequent haplotype. The haplotypes 1-2 and 3-2 have slightly protective (although not significant) effects. The presented double-dose relative risks are the $\tilde{F}_i$-parameters. For instance, $\tilde{F}_{4-2}$ represents the elevation in risk obtained by starting with a heterozygous individual having no 4-2 haplotype and replacing both haplotypes with 4-2. Since the double-dose estimate is of the same magnitude as the single dose, this would suggest that 4-2 has a dominant deleterious effect. It should be kept in mind, however, that homozygotes may be too rare to allow an accurate estimate of the dose effect pattern for all haplotypes. When evaluating significance it is clear that only haplotype 4-2 has a (borderline) significant fetal effect with a Wald p-value of 0.053. The overall likelihood ratio p-value for the MSX1 locus is 0.78, thus the effect of haplotype 4-2 is not strong enough to show up on the overall test.

Discussion

We have demonstrated the feasibility of extending the log-linear model to loci with multiple alleles, and to loci with multiple haplotypes with unknown phase. In addition to estimating fetal effects, both for single and double doses of a haplotype, most other commonly discussed effects such as the effects of maternal haplotypes, gene-environment interactions and parent-of-origin can be incorporated, and likelihood ratio tests are

Table 1 Relative risk estimates for both fetal and maternal haplotypes of the MSX1 homeoboxgene on chromosome 4. Estimates show the risk of cleft lip or palate for the child when either the child or the mother carries one (single-dose) or two (double-dose) copies of the haplotype in question. The reference level for each haplotype is heterozygotes without that haplotype. Confidence intervals (95%) are shown in parentheses. The double dose estimates of haplotypes 4-1 and 3-2 are omitted due to few homozygotes. A separate column shows the estimated population frequencies of the haplotypes. A graphical representation of the estimates for the fetal haplotypes is found in Figure 1

		Fetal haplotype relative risk		Maternal haplotype relative risk	
Haplotype	Frequency	Single-dose	Double-dose	Single-dose	Double-dose
4-1	0.03 (0.02, 0.06)	1.30 (0.63, 2.60)	–	0.87 (0.42, 1.80)	–
1-2	0.11 (0.08, 0.15)	0.73 (0.47, 1.20)	1.10 (0.36, 3.50)	1.20 (0.76, 1.90)	0.85 (0.10, 6.80)
2-2	0.28 (0.23, 0.33)	1.10 (0.76, 1.60)	1.00 (0.52, 2.00)	0.96 (0.64, 1.40)	0.93 (0.48, 1.90)
3-2	0.07 (0.04, 0.10)	0.64 (0.37, 1.10)	–	1.30 (0.76, 2.20)	–
4-2	0.51 (0.46, 0.56)	1.60 (1.00, 2.60)	1.60 (0.90, 2.90)	0.73 (0.48, 1.10)	0.84 (0.50, 1.40)

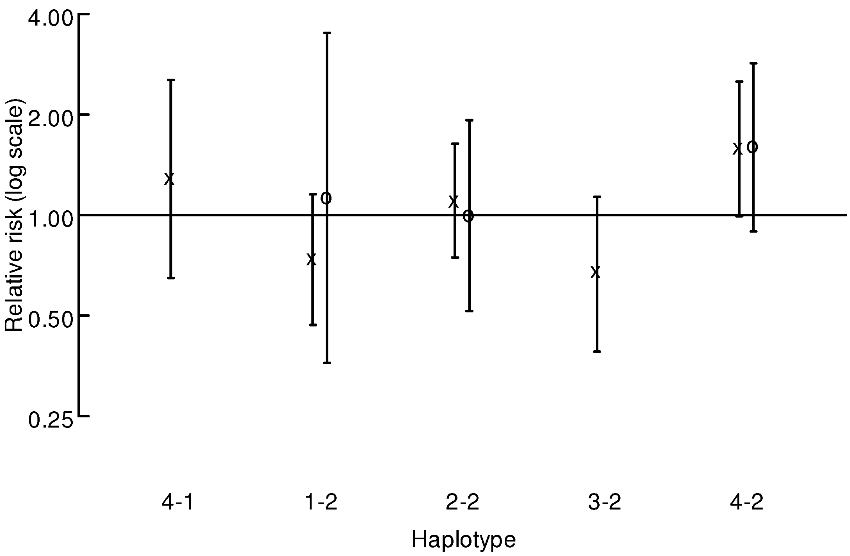

Figure 1 Relative risk estimates for the fetal haplotypes of the MSX1 homeoboxgene on chromosome 4. Estimates show the risk of cleft lip or palate for a child carrying one or two copies of the haplotype in question. The horizontal line at 1.00 marks the reference level, which for each haplotype is a heterozygote without that haplotype. Single dose estimates are marked with an "×" , double dose estimates with an "○" . The double dose estimates of haplotypes 4-1 and 3-2 are omitted due to few homozygotes. Confidence intervals (95%) are drawn as vertical lines. Numerical results for both fetal and maternal alleles are shown in Table 1.

available for nested models. As shown in the Appendix, the multiplicative structure allows for several simplifications of likelihood computations, and this makes estimation fairly straightforward even with larger-size problems. It will, however, typically require some extra programming since standard log-linear software will have problems with the size of the design matrix. The log-linear approach is particularly well suited in combination with the EM algorithm, and this makes extensions like predicting phase of phase-unknown haplotypes and accounting for missing parental information easy. In particular, the E-step can combine expectation for both unknown phase and missing parental information in the same step. These features are incorporated into the HAPLIN software.

One of the prominent features of the triad design is the possibility to control for confounding caused, for instance, by population stratification. To exploit this fully, the most common approach for triad analyses is to condition on mating type, which represents the "extreme" position that most of the relevant information should be extracted from within-family contrasts, and that across-family comparisons are more suspect due to unknown population structure (Sham & Curtis, 1995; Cordell, 2004a; Cordell *et al.* 2004b). When assuming HWE, some of this protection against population stratification is lost, but since transmitted and non-transmitted alleles are still matched the confounding effect will be of a less serious nature than in a case-control study. In addition, only a few instances are known where population stratification has really been strong enough to produce such a confounding effect (Schaid, 2002); (Wacholder *et al.* 2002). Completely ignoring the knowledge that a population is homogeneous and most likely close to HWE will waste information (Knapp *et al.* 1995), and it is not clear that the extra protection against confounding is worth the price.

It should be remarked that Hardy-Weinberg equilibrium does not have to be taken at its face value. There is information available from the triad data that may allow us to test its correctness. For instance, under the models described above, the non-transmitted allele of the father is statistically independent of the transmitted allele of the father and of the two maternal alleles. A test for this independence can be included in an analysis. Also, there is the possibility of testing for HWE, or to some extent compensating for lack of HWE by including extra parameters in the model (as described above), thus offering at least an opportunity to evaluate the impact of the HWE assumption. For the analysis of the MSX1 homeoboxgene shown in this paper there was no sign of deviations from the HWE (results not shown).

An added advantage of the triad design over, for instance, the case-control design is the ability to infer parent-of-origin for many of the transmitted alleles. This makes it possible to establish phase in many of the children directly. The children with known phase are basically those for whom the parent-of-origin can be deduced for all involved markers. This is usually the majority of the children when only few markers are involved, and thus estimating unknown phase frequencies will not be necessary. However, as more markers are added there may be a substantial number of triads for which the phase of the fetal haplotypes cannot be determined, and disregarding those triads may lead to an unacceptable loss of power. The log-linear model includes these triads in a natural fashion through the EM algorithm.

In conclusion, the proposed extension of the log-linear model approach works well for estimation of haplotype effects from case-parent triad data. Both ambiguous haplotypes and incomplete triads may be included in the analyses. Effect estimates are essential for an assessment of consistency of association with other studies, for example case-control studies. Estimation of both single-dose and double-dose effects of haplotypes is implemented in software that is available on the web.

Electronic Database Information

HAPLIN can be downloaded from http://www.uib.no/smis/gjessing/genetics/software/haplin, together with examples explaining use and interpretation of output. It runs in both S-Plus and R; the latter can be downloaded free of charge (R,2004). HAPLIN runs on most platforms, is easy to install and requires no previous knowledge of S-Plus or R. It reads data from several formats. Most of the log-linear models described in this paper are implemented.

Acknowledgements

We thank Dr. Astanand Jugessur who was responsible for the MSX1 genotyping of the oral cleft triads , and Prof. Sven Ove Samuelsen for useful comments regarding the jackknife procedure. Hilde-Gunn Bruu has contributed with valuable programming assistance, in particular for the data preparation functions in HAPLIN. The work was supported by NIH Grant No. 2R01 DE011948-04 and Research Council of Norway Grant No. 166329/V50.

Appendix

The Fully Multiplicative Model

In the diallelic situation, Weinberg *et al.* (1998) show how the model (4) can easily be implemented using standard software. As they remark, this can in principle also be done for a multiple allele situation, by setting up appropriate design matrices as in (5). For small K, in particular the diallelic situation, this is undoubtedly the easiest alternative since it also computes estimates for standard errors, and other effects such as effects of maternal genes and gene-environment interaction are effortlessly incorporated. However, as K increases the size of the design matrices increases dramatically. Since a data set in practice will be of moderate size, most cells will be empty for large K. Nevertheless, the empty cells are not *structural* zeros and thus have to be included in the likelihood. In standard software this is not easily achieved without setting up the full design matrix. For this reason, we will look at some computational simplifications when K is large. *In the following we will assume that* $K \geq 3$.

The fully multiplicative model, i.e. the model without separate double-dose parameters is simple and we will only sketch the likelihood derivation in the following. Recall the contribution (6) made by each triad type to the log-likelihood (up to additive constants). The total log-likelihood is then

$$l = \sum_{ijkl} [n_{ijkl} \log \xi_{ijkl} - \xi_{ijkl}], \tag{9}$$

where ξ_{ijkl} are the expected cell frequencies and n_{ijkl} the observed frequencies, as before. When there are no separate double-dose parameters, i.e. $R_i^\star = 1$ for all i, and no maternal effects, we write $\xi_{ijkl} = \xi \cdot p_i p_j p_k p_l \cdot R_j R_l$. Define $\lambda_i = p_i R_i$ and write $\xi_{ijkl} = p_i p_k \cdot \lambda_j \lambda_l$. Notice that estimating λ_i is equivalent to estimating R_i once p_i is estimated. We assume $\sum_i p_i = 1$, and to restrict the remaining parameters we use $\sum_i \lambda_i = 1$. Expanding the log-likelihood gives

$$l = \sum_i (n_{i\cdots} + n_{\cdot\cdot i \cdot}) \log p_i + \sum_i (n_{\cdot i \cdot\cdot} + n_{\cdots i}) \log \lambda_i + n \log \xi - \xi.$$

The estimates for p_i and λ_i can thus be obtained separately, and both correspond to simple multinomial estimates, yielding

$$\hat{p}_i = \frac{n_{i\cdots} + n_{\cdot\cdot i\cdot}}{2n} \quad \text{and} \quad \hat{\lambda}_i = \frac{n_{\cdot i\cdot\cdot} + n_{\cdots i}}{2n}, \tag{10}$$

thus the allele frequencies are estimated simply by allele counting over the non-transmitted alleles. In addition,

$$\hat{R}_i = \frac{\hat{\lambda}_i}{p_i} = \frac{n_{\cdot i\cdot\cdot} + n_{\cdots i}}{n_{i\cdots} + n_{\cdot\cdot i\cdot}},$$

i.e. the relative risks are

$$\frac{\hat{R}_j}{\hat{R}_i} = \frac{(n_{\cdot j\cdot\cdot} + n_{\cdots j})(n_{i\cdots} + n_{\cdot\cdot i\cdot})}{(n_{j\cdots} + n_{\cdot\cdot j\cdot})(n_{\cdot i\cdot\cdot} + n_{\cdots i})}.$$

Incidentally, this relative risk is equal to the odds ratio of transmitting A_j in preference to A_i in the standard HHRR (haplotype-based haplotype relative risk) model context (Terwilliger & Ott, 1992).

When including maternal effects a completely explicit solution using model (8) still exists when setting $R_i^\star = M_i^\star = 1$ for all i. Introducing new parameters $\alpha_i = p_i M_i$, $\beta_j = p_j M_j R_j$ and $\lambda_l = p_l P_l$ and the restrictions $\sum \alpha_i = \sum \beta_j = \sum \lambda_l = 1$ we can find the explicit MLEs $\hat{p}_k = n_{\cdot\cdot k\cdot}/n$, $\hat{\alpha}_i = n_{i\cdots}/n$, $\hat{\beta}_j = n_{\cdot j\cdot\cdot}/n$ and $\hat{\lambda}_l = n_{\cdots l}/n$, leading to $\hat{M}_i = \hat{\alpha}_i/\hat{p}_i = n_{i\cdots}/n_{\cdot\cdot i\cdot}$, $\hat{R}_j = n_{\cdot j\cdot\cdot}/n_{j\cdots}$ and $\hat{P}_l = n_{\cdots l}/n_{\cdot\cdot l\cdot}$. This result has a natural interpretation: the allele frequencies are estimated from the non-transmitted paternal alleles, which are unrelated to risk. The effect of maternal alleles is a contrast between alleles not transmitted from the mother and alleles not transmitted from the father. The effect of fetal genes derived from the mother is a contrast between transmitted and non-transmitted alleles from the mother, and finally the effect of fetal genes derived from the father is a contrast between transmitted and non-transmitted alleles from the father. This very simple result still requires knowledge of parent-of-origin, but that is easily overcome by application of the EM algorithm.

An EM Approach to Estimating Double-dose Effects

As seen above, when the model is purely multiplicative simple explicit estimates exist. In the following subsections we will discuss some simplifying approaches to the estimation when double-dose parameters are present. We write $\xi_{ijkl} = p_i p_k \cdot \lambda_j \lambda_l \cdot R_{jl}^\star$ with $\lambda_i = p_i R_i$ as above, and the log-likelihood is as in (9). For computational convenience we will now restrict the parameters assuming $\sum_i p_i = 1$ and $\xi = 1$. Then all of the R_i-parameters (or equivalently the λ_i-parameters) must be estimated, and so must the $R_{jl}^\star$-parameters. After expanding (9), we see that the only part containing $\mathbf{p} = (p_1, \ldots, p_k)^{\mathrm{T}}$ is

$$\sum_{ijkl} n_{ijkl} [\log(p_i) + \log(p_k)] = \sum_i (n_{i\cdots} + n_{\cdot\cdot i\cdot}) \log(p_i),$$

just as in the fully multiplicative model (10). The remaining part of the log-likelihood can then be summed over i and k to obtain

$$\sum_{jl}\left[n_{\cdot j\cdot l}\log\left(\lambda_j\lambda_l\cdot R^\star_{jl}\right)-\lambda_j\lambda_l\cdot R^\star_{jl}\right], \tag{11}$$

still disregarding additive constants. This is the log-likelihood of a Poisson model with cell expected values $\lambda_j\lambda_l R^\star_{jl}$ and observed cell frequencies $n_{\cdot j\cdot l}$. By differentiating with respect to the $R^\star_j$-parameters it is seen that $\hat{R}^\star_j = n_{\cdot j\cdot j}/\hat{\lambda}_j^2$ is the MLE for $R^\star_j$ once $\hat{\lambda}_j$ has been found. This is, in fact, the value of $\hat{R}^\star_j$ that makes the expected values match the observed perfectly in all categories with homozygous children ($j = l$). Entering $\hat{R}^\star_j$ in (11) yields the remaining

$$\sum_{j\neq l}[n_{\cdot j\cdot l}\log(\lambda_j\lambda_l)-\lambda_j\lambda_l] \tag{12}$$

of the log-likelihood to be used to estimate the λ_is. We will now show how this can be maximized as a part of the EM algorithm; in the next subsection we will show that a near-explicit solution can also be found. Observe that (12) is just the log-likelihood of a fully multiplicative model, only the diagonal elements ($j = l$) are removed. If the diagonal elements were included, the MLE would be attained at $\hat{\lambda}_j = (n_{\cdot j\cdot\cdot}+n_{\cdot\cdot\cdot j})/\sqrt{n}$. In addition, if the parameters λ_j were known the expected numbers in the diagonal cells in a fully multiplicative model would be λ_j^2. Thus, the actual MLEs for λ_j from (12) can be estimated using an EM algorithm. We completely disregard the actual observed values $n_{\cdot j\cdot j}$ at the diagonal and say that the full data for the EM algorithm consist of the observed off-diagonal values $n_{\cdot j\cdot l}$ together with unobserved frequencies m_j when $j = l$. The M step is then to compute $\hat{\lambda}_j = (n_{\cdot j\cdot\cdot}+n_{\cdot\cdot\cdot j})/\sqrt{n}$ with $n_{\cdot j\cdot j}$ replaced by m_j for all j. The E step consists of updating the expected values of the frequencies conditional on the observed data, which just amounts to updating m_j to the value $\hat{\lambda}_j^2$, using the current updates of the parameter estimates $\hat{\lambda}_j^2$.

In summary, the EM algorithm would thus be to first compute $\mathbf{p}$ from the non-transmitted alleles. Next, estimate λ as in the fully multiplicative model. Then, replace the number of homozygous children by their expected values λ_j^2. The step computing λ is then repeated, without re-estimating $\mathbf{p}$, and the number of homozygotes expected from the multiplicative model again replace the previous expected values. When convergence is achieved and the estimates $\hat{\mathbf{p}}$ and $\hat{\lambda}$ are known, we compute $\hat{R}_j = \hat{\lambda}_j/\hat{p}_j$ and $\hat{R}^\star_j = n_{\cdot j\cdot j}/\hat{\lambda}_j^2$, where $n_{\cdot j\cdot j}$ are the original (observed) frequencies.

The advantage of estimating the parameters in the EM framework is that it can be incorporated with the EM algorithm necessary to reconstruct missing data or missing haplotype information. We remark that similar use of the EM algorithm in tables with missing diagonals is described, for instance, in Morgan & Titterington 1977.

An Explicit Solution for the Double-dose Model

Although the EM approach described above is probably the easiest to implement, for the double-dose model without maternal effects there is a solution to the maximum likelihood estimation which requires no iterations except to estimate a single parameter from a well-behaved one-dimensional equation. Once this is done, all other parameters can be computed from this single estimate. We will give the details of the computations below since this amounts to an almost completely explicit solution which is computationally very fast and places only minimal requirements on memory and other resources. A similar derivation for non-symmetric tables is found in (Wagner, 1970).

As shown above, the allele frequencies $\mathbf{p}$ are estimated explicitly from the non-transmitted alleles, and the R_i and $R^\star_i$ can always be computed once the λ_i have been estimated. To estimate λ_i, consider again the part (12) of the likelihood needed to estimate λ_i. Differentiating with respect to λ_i we obtain the equation

$$\hat{\lambda}_i^2-\hat{B}\hat{\lambda}_i+\gamma_i/4=0$$

for $\hat{\lambda}_i$, where we define $B=\sum_j\lambda_j$, $\hat{B}=\sum_j\hat{\lambda}_j$ and $\gamma_i = 2(n_{\cdot i\cdot\cdot}+n_{\cdot\cdot\cdot i}-2n_{\cdot i\cdot i})$, i.e. two times the number of heterozygous children with one A_i-allele. Notice that *if $\hat{B}$ were known* this is a second-order equation in $\hat{\lambda}_i$ and it can be solved by the standard formula

$$\hat{\lambda}_i=\frac{1}{2}\hat{B}\left\{1-s_i\sqrt{1-\gamma_i/\hat{B}^2}\right\}$$

for all $i = 1, 2, \ldots, K$. Notice that the sign $s_i = \pm 1$ depends on i . It is not entirely obvious how to pick the correct sign, but we will derive a simple rule for this below. For the time being, define the functions $f_i(x) = s_i\sqrt{1 - \gamma_i/x^2}$, so that the equation becomes

$$\hat{\lambda}_i = \frac{1}{2}\hat{B}(1 - f_i(\hat{B})). \tag{13}$$

Thus, the only thing that needs to be computed numerically is $\hat{B}$. Once this is done, $\hat{\lambda}_i$ can be computed from (13). To derive an equation for $\hat{B}$, sum both sides of (13) over $i = 1, \ldots, K$ to get

$$K - 2 = \sum_i f_i(\hat{B}), \tag{14}$$

which is easily solved numerically as an equation of $\hat{B}$. When $\hat{B}$ has been obtained, all other parameter estimates are computed as described above.

The only remaining difficulty is how to choose the sign s_i. The derivation is somewhat involved, but the resulting rule is easy to implement. Choose m such that $\gamma_m > \gamma_i$ for all $i \neq m$ and set $M = \sqrt{\gamma_m} = \max\{\sqrt{\gamma_i}; i = 1, 2, \ldots, K\}$. Compute the sum $S = \sum_i f_i(M)$. The rule is: If $S < K - 2$, choose $s_i = +1$ for all i. If $S > K - 2$ then choose $s_m = -1$ but $s_i = +1$ for all $i \neq m$.

The argument for this rule is as follows. Assume first that all $s_i = +1$. Then the functions $f_i(x)$ are all increasing in x, they are all defined when $x \geq M$ and $\lim_{x\to\infty} f_i(x) = 1$. Thus, $S = \sum_i f_i(M)$ is the smallest value attained by $\sum_i f_i(x)$, and $\lim_{x\to\infty} \sum_i f_i(x) = K$. Hence, when $S < K - 2$ equation (14) must have a unique positive solution $\hat{B} > M$. If $S > K - 2$ it will *not* have a solution. However, when choosing $s_m = -1$ but keeping the other signs positive it can be seen that this will produce a solution. Since each $f_i(x)$ is decreasing as a function of γ_i it can be deduced that any other combination of signs will make the sum on the right hand side of (14) always less than $K - 2$, thus never yielding a solution.

References

Cheng, R., Ma, J. Z., Wright, F. A., Lin, S., Gao, X., Wang, D., Elston, R. C. & Li, M. D. (2003) Nonparametric disequilibrium mapping of functional sites using haplotypes of multiple tightly linked single-nucleotide polymorphism markers. *Genetics* **164**, 1175–1187.

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

Cordell, H. J. (2004) Properties of case/pseudocontrol analysis for genetic association studies: effects of recombination, ascertainment, and multiple affected offspring. *Genet Epidemiol* **26**(3):186–205.

Cordell, H. J., Barratt, B. J. & Clayton, D. G. (2004) Case/pseudocontrol analysis in genetic association studies: a unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects. *Genet Epidemiol* **26**(3):167–185.

Dudbridge, F. (2003) Pedigree Disequilibrium Tests for Multilocus Haplotypes. *Genet Epidemiol* **25**, 115–121.

Efron, B. & Tibshirani, R. J. (1993) *An Introduction to the Bootstrap*. Chapman & Hall, New York.

Falk, C. T. & Rubinstein, P. (1987) Haplotype relative risks - An easy reliable way to construct a proper control sample for risk calculations. *Annals of Human Genetics* **51**, 227–233.

Horvath, S., Xu, X., Lake, S. L., Silverman, E. K., Weiss, S. T. & Laird, N. M. (2004) Family-based tests for associating haplotypes with general phenotype data: application to asthma genetics. *Genet Epidemiol* **26**, 61–69.

International human genome sequencing consortium (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.

International SNP map working group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933.

Jugessur, A., Lie, R. T., Wilcox, A. J., Murray, J. C., Taylor, J. A., Saugstad, O. D., Vindenes, H. A. & Abyholm, F. (2003) Variants of developmental genes (TGFA, TGFB3, and MSX1) and their associations with orofacial clefts: a case-parent triad analysis. *Genet Epidemiol* **24**, 230–9.

Khoury, M. J. (1994) Case-parental control method in the search for disease-susceptibility genes. *Am J Hum Genet* **55**, 414–415.

Knapp, M., Wassmer, G. & Baur, M. P. (1995) The relative efficiency of the Hardy-Weinberg equilibrium-likelihood and the conditional on parental genotype-likelihood methods for candidate gene association studies. *Am J Hum Genet* **57**(6):1476–1485.

Laird, N. M., Horvath, S. & Xu, X. (2000) Implementing a unified approach to family-based tests of association. *Genet Epidemiol* **19** (Suppl1):S36–S42.

Lee, W.-C. (2003) Searching for disease-susceptibility loci by testing for Hardy-Weinberg disequilibrium in a gene bank of affected individuals. *Am J Epidemiol* **158**(5):397–400.

Lidral, A. C., Murray, J. C., Buetow, K. H., Basart, A. M., Schearer, H., Shiang, R., Naval, A. *et al.* (1997) Studies of the candidate genes TGFB2, MSX1, TGFA, and TGFB3 in the etiology of cleft lip and palate in the Philippines. *Cleft Palate Craniofac J* **34**, 1–6.

Lidral, A. C., Romitti, P. A., Basart A. M., Doetschman, T., Leysens, N. J., Daack-Hirsch, S., Semina, E. V. *et al.* (1998) Association of MSX1 and TGFB3 with nonsyndromic clefting in humans. *Am J Hum Genet* **63**, 557–68.

Mitchell, L. E. (2000) Relationship between case-control studies and the transmission/disequilibrium test. *Genet Epidemiol* **19**, 193–201.

Morgan, B. J. T. & Titterington, D. M. (1977) A comparison of iterative methods for obtaining maximum likelihood estimates in contingency tables with a missing diagonal. *Biometrika* **64**(2):265–9.

Editorial (1999) Freely Associating. *Nature Genetics* **22**(1):1–2.

R Development Core Team (2004) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Self, S., Longton, G., Kopecky, K. & Liang, K. (1991) On estimating HLA-disease association with application to a study of aplastic anemia. *Biometrics* **47**, 53–61.

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

Schaid, D. J. (2002) Disease-marker association. In: Biostatistical Genetics and Genetic Epidemiology (eds. R. Elston, J. Olson and L. Palmer), John Wiley & Sons, New York.

Sham, P. C. & Curtis, D. (1995) An extended transmission/disequilibrium test (TDT) for multiallele marker loci. *Annals of Human Genetics* **59**, 323–336.

Sinsheimer, J. S., Blangero, J. & Lange, K. (2000) Gamete-Competition Models. *American Journal of Human Genetics* **66**, 1168–1172.

Sinsheimer, J. S., Palmer, C. G. S. & Woodward, J. A. (2003) The maternal-fetal genotype incompatiblity test: detecting genotype combinations that increase risk for disease. *Genet Epidemiol* **24**, 1–13.

Sorensen, D. & Gianola, D. (2002) *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*. Springer Verlag, New York.

Spielman, R. S., McGinnis, R. E. & Ewens, W. J. (1993) Transmission test for linkage disequilibrium: the insuline gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics* **52**, 506–516.

Terwilliger, J. & Ott, J. (1992) A haplotype based 'haplotype relative risk' approach to detecting allelic associations. *Hum Hered* **42**, 337–346.

Thomas, D. C. (2000) Case-parent design for gene-environment interaction by Schaid. *Genet Epidemiol* **19**, 461–463.

Umbach, D. M. & Weinberg, C. R. (2000) The use of case-parent triads to study joint effects of genotype and exposure. *Am J Hum Genet* **66**, 251–261.

Wacholder, S., Rothman, N. & Caporaso, N. (2002) Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. *Cancer Epidemiol. Biomarkers Prev*. **11**, 513–520.

Wagner, S. S. (1970) The maximum-likelihood estimate for contingency tables with zero diagonal. *JASA* **65**(331): 1362–1383.

Weinberg, C. R. (1999*a*) Allowing for missing parents in genetic studies of case-parent triads. *Am J Hum Genet* **64**, 1186–1193.

Weinberg, C. R. (1999*b*) Methods for detection of parent-of-origin effects in genetic studies of case-parent triads. *Am J Hum Genet* **65**, 229–235.

Weinberg, C. R. & Umbach, D. M. (2000) Choosing a retrospective design to assess joint genetic and environmental contributions to risk. *Am J Epidemiol* **152**(3):197–203.

Weinberg, C. R., Wilcox, A. J. & Lie, R. T. (1998) A log-linear approach to case-parent-triad data: assessing effects of disease genes that act directly or through maternal effects and that may be subject to parental imprinting. *Am J Hum Genet* **62**, 969–978.

Wilcox, A. J., Weinberg, C. R. & Lie, R. T. (1998) Distinguishing the effects of maternal and offspring genes through studies of "case-parent triads." *Am J Epidemiol* **148**, 893–901.

Zhao, H. (2000) Family-based association studies. *Statistical Methods in Medical Research* **9**, 563–587.

Zhao, H., Zhang, S., Merikangas, K. R., Trixler, M., Wildenauer, D. B., Sun, F. & Kidd, K. K. (2000) Transmission/disequilibrium tests using multiple tightly linked markers. *Am J Hum Genet* **67**, 936–946

Received: 30 October 2004
Accepted: 25 July 2005