A General Likelihood Approach to Trait-Based Multipoint Linkage Analysis in Large Groups of Half-Sibs and Super Sisters

Norbert Reinsch

Institut für Tierzucht und Tierhaltung, Christian-Albrechts Universität zu Kiel, D-24098 Kiel, Germany

Manuscript received January 11, 2002
Accepted for publication May 28, 2002

ABSTRACT

The idea of trait-based linkage analysis in half-sibs is extended by comparing the frequency of parental marker haplotypes in animals with different phenotypes. This article first presents the likelihood of observing different classes of paternal haplotypes in a half-sib family, where only family members of a certain phenotype (e.g., affected) are genotyped and are fully informative. The likelihood function is then generalized to multiple phenotypic categories. A linear predictor allows for discontinuous as well as for continuous phenotypes and other explanatory variables. Finally, how to incorporate not fully informative offspring and how to analyze super sister families are shown. Maximum-likelihood estimates of all parameters can be found by a Newton-Raphson algorithm, which mimics an iteratively weighted least-squares procedure. The method allows for any multilocus feasible mapping function and, among others, for situations with selective or nonselective genotyping, single or multiple traits, and continuous or categorical traits. No parameters are required to describe the mode of inheritance and the method copes with virtually any family size. Fields of applications are therefore mapping experiments in species with a high reproductive capacity, such as cattle, pigs, horses, honey bees, trees, and fish.

The term “trait-based” analysis was coined by Lebowitz et al. (1987) for a certain type of linkage analysis. The basic idea is to compare marker genotypes or marker allele frequencies of different phenotypes as opposed to a comparison of phenotypic means of different marker genotypes. The latter type of analysis has been termed “marker based” (Lebowitz et al. 1987) and includes well-known techniques such as, e.g., maximum-likelihood or regression interval mapping (Lander and Botstein 1989; Haley and Knott 1992; Martinez and Curnow 1992).

Lebowitz et al. (1987) stated that trait-based analysis may be a useful alternative when interest is centered on a single quantitative trait only and marker-based analysis is not applicable as for polygenic resistance traits, where only a part of the population survives exposure to the stressor. Henshall and Goddard (1999) showed that the probability of a paternal $Q$ allele in daughters of a sire heterozygous $Qq$ at a quantitative trait locus (QTL) shows a logistic relationship with the trait influenced by the QTL if the error within QTL genotype is normally distributed. They used logistic regression for the estimation of QTL effects in selectively genotyped half-sib families, treating marker genotype of progeny (first vs. second paternal marker allele) as the binary response variable and phenotype as the independent variable. Estimates for the QTL effect were unbiased by selecting animals for genotyping from the extreme tails of the distribution of phenotypes. Henshall and Goddard (1999) also demonstrated that multiple-trait QTL detection in half-sib families is a relatively simple matter of performing multivariate logistic regression, no matter if selective genotyping has been applied or not.

With half-sib experiments, in treating genotype of the offspring as response, the binary response variable (paternal $Q$ or paternal $q$) can be observed directly, provided there is complete linkage with the marker under consideration and all progeny are fully informative. The numbers of paternal $Q$ and $q$ alleles cannot be directly observed and counted in situations with recombination, if the marker allele inherited from the sire cannot be determined unequivocally for a part of the sibship or if a pair of flanking markers is available. In these cases it will often be possible to resolve the marker genotypes of the offspring into parental marker haplotypes (or sets of possible parental haplotypes compatible with the marker data) and these haplotypes may alternatively be taken as observations. The distributions of these observations are, however, no longer binomial, but, in the general case, a mixture of different multinomial and binomial distributions.

This article deals first with the likelihood to observe different classes of paternal haplotypes in a half-sib family, where only family members of a certain phenotype (e.g., affected) are genotyped and are fully informative for either two flanking markers or a single marker. The likelihood function is then generalized to multiple phenotypic categories. A linear predictor on the log-scale allows for discontinuous as well as for continuous pheno-
types and other explanatory variables such as, e.g., sex or age of onset. Finally it is shown how not fully informative offspring can be incorporated, how super sister families with a degree of relationship of 0.75 can be analyzed, and that any multilocus feasible mapping function can be used. Maximum-likelihood estimates of all parameters can be found by a Newton-Raphson algorithm, which can be formulated as an iteratively weighted least-squares procedure. Application to various experimental designs and hypothesis testing in different situations is discussed.

THEORY

Half-sib setting: In livestock species such as, e.g., dairy cattle, large paternal half-sib families are available for gene mapping purposes. For a dichotomous trait with unknown mode of inheritance a simple Mendelian analysis is not possible. However, the probability of paternal half-sibs to share a certain paternal chromosome segment given the marker data and phenotypes may be evaluated.

The likelihood function for a single half-sib family with \( N \) affected offspring is then

\[
L = \prod_{i=1}^{N} \prod_{j=1}^{k} h_{ij}^{n_{ij}}
\]

and the log-likelihood \( l \) is

\[
l = \sum_{i=1}^{N} \sum_{j=1}^{k} \delta_{ij} \ln h_{ij},
\]

where \( i \) is the class of the paternal haplotype of an affected family member, \( k = 6 \) if either a single marker or a marker pair is available for a certain offspring, and \( \delta_{ij} \) is a 0–1 indicator variable with value 1 if the paternal haplotype of an affected offspring is of type \( i \) and a value of 0 otherwise.

The maximum of the log-likelihood at a certain map position can be found by a Newton-Raphson algorithm using the first and second derivatives with respect to \( c \):

\[
\frac{\partial l}{\partial c} = \sum_{i=1}^{N} \sum_{j=1}^{k} \delta_{ij} \frac{\partial \ln h_{ij}}{\partial c} = \sum_{i=1}^{N} \sum_{j=1}^{k} \delta_{ij} \frac{1}{h_{ij}} \frac{\partial h_{ij}}{\partial c}
\]

and

\[
\frac{\partial^{2} l}{\partial^{2} c} = \sum_{i=1}^{N} \sum_{j=1}^{k} \delta_{ij} \frac{\partial^{2} \ln h_{ij}}{\partial^{2} c}. \tag{5}
\]

There are again six different possibilities of how an affected family member can contribute to the likelihood and to the first and second derivatives of the log-likelihood function, depending on its paternal haplotype. The different types of contributions to the likelihood and to the first and second derivatives can be expressed in a more general form as functions of \( c \) and two quantities \( a \) and \( b \), which depend on the recombination rates between the susceptibility locus and the markers.

\[
h_{i} = \frac{h_{i} + a c}{1 + c}
\]

\[
\frac{\partial \ln h_{i}}{\partial c} = \frac{a_{i} - h_{i}}{(b_{i} + a c)(1 + c)}
\]

\[
\frac{\partial^{2} \ln h_{i}}{\partial^{2} c} = \frac{-(a_{i} - h_{i})(a_{i} + b_{i} + 2 a c)}{(b_{i} + a c)^{2}(1 + c)^{2}}. \tag{8}
\]

The different \( a \) and \( b \) values are shown in Table 1.

Multiple phenotypic categories: The analysis is not
necessarily restricted to a single phenotypic category and the inclusion of unaffected animals may be of interest. The probabilities of having inherited one of the six classes of paternal haplotypes if an animal is unaffected are shown in the first column of Table 2. These formulas can be rewritten by introducing the ratio \( c_k = (1 - \phi_i) / (1 - p) \) and replacing \( (1 - \phi_i) \) with \( (1 - p) c_k \). This yields formulas with the same structure as for affected animals, but with \( c \) replaced by \( c_k \), as can be seen in detail in the second column of Table 2. It is obvious that \( c_k \) is also equal to 1 under the null hypothesis. The log-likelihood function \( l \) for observations from both categories is composed of contributions of the affected animals and contributions of the unaffected animals,

\[
l = \sum_{s=1}^{N} \sum_{i=1}^{k} \delta_i \ln [h_i(\epsilon)] + \sum_{s=1}^{N} \sum_{i=1}^{k} \delta_i \ln [h_i(\epsilon_k)],
\]

where \( N \) and \( N_r \) are the numbers of affected and unaffected animals in the half-sib family, respectively.

More generally, if there are \( r \) discrete phenotypic categories, the number of half-sibs in each category is \( N_r \), and a parameter \( \epsilon_i \) is defined for each category in a similar manner as shown above (with \( \epsilon_i = c \)), the log-likelihood can be written as

\[
l = \sum_{s=1}^{N} \sum_{i=1}^{k} \delta_i \ln [h_i(\epsilon)] + \sum_{s=1}^{N} \sum_{i=1}^{k} \delta_i \ln [h_i(\epsilon_k)],
\]

where

\[
l = \frac{1}{2} \log \left[ \frac{1 - \theta_1}{1 - \theta_2} \right] + \frac{1}{2} \log \left( \frac{1 - \theta_1}{1 - \theta_2} \right)
\]

\[
\theta_1 = \frac{1 - \phi_i}{1 - p} \text{ and } \theta_2 = \frac{p}{1 - \phi_i}
\]

Log odds and linear predictor: Another method of observing the problem is to consider the ratio of the two probabilities \( p(\text{paternal } D | \text{markers, phenotype}) \) and \( p(\text{paternal } d | \text{markers, phenotype}) \) for affected half-sibs. If an affected individual has paternal haplotype 1-1 these probabilities are

\[
p(D|1-1, \text{affected}) = \frac{1}{2} \log \left[ \frac{1 - \theta_1}{1 - \theta_2} \right] + \frac{1}{2} \log \left( \frac{1 - \theta_1}{1 - \theta_2} \right) + \frac{1}{2} \log \left[ \frac{1 - \phi_i}{1 - p} \right]
\]

and

\[
p(d|1-1, \text{affected}) = \frac{1}{2} \log \left[ \frac{1 - \theta_1}{1 - \theta_2} \right] + \frac{1}{2} \log \left( \frac{1 - \theta_1}{1 - \theta_2} \right) + \frac{1}{2} \log \left[ \frac{1 - \phi_i}{1 - p} \right]
\]

The corresponding ratio (odds) in favor of \( D \) is then

\[
\text{Odds} = \frac{p(D|1-1, \text{affected})}{p(d|1-1, \text{affected})}
\]

\[
= \frac{\frac{1}{2} \log \left[ \frac{1 - \theta_1}{1 - \theta_2} \right] + \frac{1}{2} \log \left( \frac{1 - \theta_1}{1 - \theta_2} \right) + \frac{1}{2} \log \left[ \frac{1 - \phi_i}{1 - p} \right]}{\frac{1}{2} \log \left[ \frac{1 - \theta_1}{1 - \theta_2} \right] + \frac{1}{2} \log \left( \frac{1 - \theta_1}{1 - \theta_2} \right) + \frac{1}{2} \log \left[ \frac{1 - \phi_i}{1 - p} \right]}
\]

This method provides a way to assess the likelihood of \( D \) being present in the sample, given the observed phenotype.

**Figure 1.**—Probability tree showing the coinheritance of alleles at a trait locus \((d, D)\) and two flanking markers in a paternal half-sib family. Probabilities can be seen at the branches of the tree: \( \theta_1 \) and \( \theta_2 \) are recombination rates between trait locus and markers, and \( p_k \) and \( r_k \) are disease probabilities for family members with a paternal \( d \) or \( D \) allele, respectively. Affected status is denoted by \( a \) (affected) and \( u \) (unaffected).
TABLE 1
Probabilities of observing different classes of paternal haplotypes in affected half-sibs

<table>
<thead>
<tr>
<th>Paternal haplotype</th>
<th>Index i</th>
<th>Probability ( h_i ) of paternal haplotype if individual is affected</th>
<th>( a_i )</th>
<th>( b_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1</td>
<td>1</td>
<td>( h_1 = \frac{(1 - \theta_1)(1 - \theta_2) + \theta_0 \theta_2 c}{(1 + c)} )</td>
<td>( \theta_0 )</td>
<td>((1 - \theta_1)(1 - \theta_2))</td>
</tr>
<tr>
<td>1-2</td>
<td>2</td>
<td>( h_2 = \frac{(1 - \theta_2) \theta_0 + \theta_1 (1 - \theta_2) c}{(1 + c)} )</td>
<td>( \theta_1 )</td>
<td>((1 - \theta_2)\theta_2 )</td>
</tr>
<tr>
<td>2-1</td>
<td>3</td>
<td>( h_3 = \frac{\theta_1 (1 - \theta_2) + (1 - \theta_0) \theta_2 c}{(1 + c)} )</td>
<td>( (1 - \theta_1) \theta_2 )</td>
<td>( \theta_1 (1 - \theta_2) )</td>
</tr>
<tr>
<td>2-2</td>
<td>4</td>
<td>( h_4 = \frac{\theta_0 \theta_2 + (1 - \theta_1)(1 - \theta_2) c}{(1 + c)} )</td>
<td>( (1 - \theta_1)(1 - \theta_2) )</td>
<td>( \theta_0 )</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>( h_5 = \frac{(1 - \theta_0) + \theta_0 c}{(1 + c)} )</td>
<td>( 0 )</td>
<td>((1 - 0))</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>( h_6 = \frac{\theta + (1 - \theta) c}{(1 + c)} )</td>
<td>( (1 - 0) )</td>
<td>( 0 )</td>
</tr>
</tbody>
</table>

The probabilities of observing different classes of paternal haplotypes in affected half-sibs are expressed as functions of the parameter \( c \) and the recombination rates \( \theta_0 \) and \( \theta_2 \) between the trait locus and the first and second flanking markers, respectively.

At first we consider derivatives of the single contributions to the log-likelihood with respect to \( \eta_i \) as

\[
\frac{\partial l}{\partial \eta_i} = \sum_{j=1}^{k} \frac{\partial l}{\partial \eta_j} \frac{\partial \eta_j}{\partial \eta_i} = \sum_{j=1}^{k} \left[ \frac{a_i - b_i}{b_i + a_i \exp(\eta_j)} \right] \cdot [1 + \exp(\eta_j)]
\]

where summation is over the \( k \) \((k = 6)\) different kinds of paternal haplotypes and \( \eta_i \) denotes the number of individuals with identical explanatory variables and identical paternal haplotype. The first derivative of the log-likelihood with respect to \( \beta \) is then a vector, which itself can be expressed as a sum of vectors.

\[
\frac{\partial l}{\partial \beta} = \sum_{j=1}^{k} \frac{\partial l}{\partial \eta_j} \frac{\partial \eta_j}{\partial \beta} = \sum_{j=1}^{k} \frac{\partial l}{\partial \eta_j} \frac{\partial \eta_j}{\partial \beta} = X' \frac{\partial l}{\partial \eta_j}
\]

where \( X' \) is a matrix with rows \( x_j' \) and can be interpreted as a design matrix. The second derivative of the likelihood with respect to \( \beta \) gives a matrix:

\[
\frac{\partial^2 l}{\partial \beta \partial \beta} = \sum_{j=1}^{k} \left[ \frac{\partial^2 l}{\partial \eta_j \partial \beta} \frac{\partial \eta_j}{\partial \beta} + \frac{\partial^2 \eta_j}{\partial \beta \partial \beta} \frac{\partial l}{\partial \eta_j} \right]
\]

\[
= \sum_{j=1}^{k} \frac{\partial^2 l}{\partial \eta_j \partial \beta} x_j x_j' + X' \frac{\partial^2 \eta_j}{\partial \beta \partial \beta} X
\]

Finally the log odds are

\[
\log \text{odds} = \ln \frac{a_i \theta_2}{b_i (1 - \theta_0) (1 - \theta_2)} + \ln c.
\]

The log odds for each kind of paternal haplotype can be written as

\[
\eta_j = \ln a_i - \ln b_i + \ln c_j,
\]

where \( \ln a_i - \ln b_i \) is known \textit{a priori} for a given map position and is defined as shown in Table 1. In the context of generalized linear models such a term is called an “offset” (McCullagh and Nelder 1989). The logarithm of the unknown quantity \( c_j \) depends on the phenotypic category and possibly on other explanatory variables like, e.g., sex of the progeny. The logarithm of \( c_j \) may therefore be replaced by the linear predictor \( \eta_j = x_j' \beta \) and the outcome is

\[
\eta_j = \ln a_i - \ln b_i + \eta_j.
\]

Estimation of \( \beta \): As already mentioned the log-likelihood is a sum of \( r \) contributions of single animals or, equivalently, contributions of \( r \) groups of animals sharing the same set of explanatory variables (Equation 10). Each contribution has index \( j \).

\[
l = \sum_{j=1}^{r} l(\eta_j)
\]

\[
\eta_j = x_j' \beta.
\]
The maximum-likelihood estimates of $\beta$ can therefore be computed by iteratively solving

$$\hat{\beta}_{n+1} = \hat{\beta}_n + [X'W_nX]^{-1}X'U_n$$

(27)

until convergence is reached ($n$ denotes iteration number). $W_n$ is a $r \times r$ diagonal matrix of second derivatives and $U_n$ a $r \times 1$ vector of first derivatives $u_i$ with respect to $\eta_i$. The diagonal elements $w_{ij}$ of the $W$ matrix can be calculated as

$$\frac{\partial^2 l}{\partial \eta_i} = \sum_{j=1}^r w_{ij}$$

(28)

with values for $\xi$ from the current iteration. Multiplying the iterative iteration equation with $X'W_nX$ gives

$$(X'W_nX)\hat{\beta}_{n+1} = (X'W_nX)\hat{\beta}_n + X'U_n = X'(W_nX\hat{\beta}_n + U_n)$$

(29)

and $\beta$ can be estimated by iterating on the equations

$$\hat{\beta}_{n+1} = (X'W_nX)^{-1}X'Z_n,$$

(30)

where the elements of the vector $Z_n$ are artificial observations

$$w_{ij} \cdot \hat{\eta}_i + u_i,$$

(31)

which have to be updated in each iteration. These estimation equations have the same structure as those used for generalized linear models (McCullagh and Nelder 1989); however, the elements of $W_n$ and $Z_n$ are computed in a quite different way without using the predicted or observed frequencies of the response variable.

**Extension to partially informative progeny:** So far only the best scenario has been considered, where all progeny are fully informative. As an example for the treatment of partially informative progeny one member of the above-mentioned paternal half-sib family may have genotypes 1, 1 and 1, 2 at the first and second marker locus, respectively. Its paternal haplotype is therefore either 1-1 or 1-2. If the paternal haplotype is 1-1 then the corresponding maternal haplotype must be 1-2. Assuming linkage equilibrium between markers, the probability of observing such an individual is $h_{f_{11}\overline{f}_{22}}$, and the alternative of an individual with paternal haplotype 1-2 and 1-1 on the maternal chromosome has a probability of $h_{f_{12}\overline{f}_{21}}$, where $f_{11}$, $f_{12}$, and $f_{22}$ are the population frequencies of the first allele at the first marker and of alleles 1 and 2 at the second marker, respectively. After defining

$$w_1 = f_{11}f_{22}$$

(32) and

$$w_2 = f_{11}f_{21}$$

(33)

we can express the probability of observing either a
paternal 1-1 or 1-2 haplotype given the affected genotype as
\[ p(1\text{-}1\text{ or}\ 1\text{-}2|\text{affected}) = w_1h_1 + w_2h_2 \] (34)
\[ = \frac{w_1(h_1 + a_1)}{1 + c} + \frac{w_2(h_2 + a_2)}{1 + c} \] (35)
\[ = \frac{(w_1h_1 + w_2h_2) + (w_1a_1 + w_2a_2)c}{1 + c} \] (36)
\[ = b^* + a^*c \]
\[ (1 + c) = h^*. \] (37)

The contribution \( h^* \) to the likelihood from the noninformative progeny of our example can thus be written in a manner similar to the contributions of the fully informative family members from Table 1. In general, once the problem of computing adequate weights and hence adequate values for \( b^* \) and \( a^* \) has been solved, these values can be used in the computation of the loglikelihood and its first and second derivatives as described above.

In paternal half-sibs with known haplotypes for the sire and ungenotyped dams randomly chosen from the population the weights are functions of the recombination rates between the markers and the population frequencies of the marker alleles. For a partially informative half-sib all possible haplotypes can be enumerated. On each of these possible haplotypes the two markers flanking the trait locus can be treated as they are in fully informative offspring. One of the \( h_i \) values given in Table 1 must be appropriate, i.e., \( h_1, h_2, h_5, \) or \( h_6 \) for the paternal allele combinations 1-1, 1-2, 2-1, or 2-2, respectively. The matching \( h_i \) value for a certain member of the set of all possible haplotypes is denoted by \( h_{\text{g}} \).

When considering several adjacent markers on a chromosome at a time one can express the probability of observing a certain haplotype with index \( g \) as
\[ h_{\text{g}} \cdot b_{\text{g}} \cdot \prod_{w=1}^{M} f_{\text{w},v} = h_{\text{g}} \left( \prod_{w=1}^{M} (1 - \theta_w) (1 - \theta_{w}) \theta_{w}^{a_w} \right) \prod_{w=1}^{M} f_{\text{w},v}, \] (38)

where \( M \) is the number of markers on the chromosome, \( \theta_w \) is the recombination frequency between markers with index \( m \) and \( m + 1, \delta_w \) is an indicator variable with a value of zero if the alleles at the marker loci \( m \) and \( m + 1 \) are jointly inherited from the sire without recombination and a value of one in the case of a recombination, \( q \) and \( q + 1 \) are the indices of the markers flanking the (postulated) position of the putative trait locus, and \( f_{\text{w},v} \) is the frequency of the maternal allele \( v \) at marker locus \( m \). The quantity \( b_{\text{g}} \) is the probability for the paternal marker alleles left and right from the flanking interval multiplied by the probability for the maternal marker alleles at all marker loci. For each of the possible haplotypes (their total number is \( G \)) a weight \( w_{\text{g}} \) can be computed:
\[ w_{\text{g}} = b_{\text{g}} \left( \prod_{w=1}^{M} f_{\text{w},v} \right) \] (39)

The values \( a_{\text{a}} \) and \( b_{\text{g}} \) corresponding to \( h_{\text{g}} \) give two values, namely
\[ a_{\text{a}} = \sum_{z=1}^{G} w_z a_{\text{z}} \] (40)
and
\[ b_{\text{g}} = \sum_{z=1}^{G} w_z b_{\text{z}} \] (41)

for each half-sib as input for the estimation procedure at a certain map position.

Note that the recombination rate \( \theta_w \) between the markers closest to the trait locus does not contribute to the first product in (38), because it is assumed that the trait locus is between two markers. If the trait locus is exactly at a marker, then the recombination rates of all marker intervals have to be used and \( b_{\text{g}} \) becomes
\[ b_{\text{g}} = \left( \prod_{w=1}^{M} (1 - \theta_w) \theta_{w}^{a_w} \right). \] (42)

Since the trait locus coincides with a marker locus, \( h_{\text{g}} \) is either \( h_5 \) or \( h_6 \) (Table 1). Both formulas (38) and (42) assume Haldane’s mapping function and linkage equilibrium between markers and the trait locus. The linkage group of markers, which is considered simultaneously, reaches from the closest fully informative marker on one side of the trait locus to the closest fully informative marker on the other side. If no fully informative marker is available on either side of the trait locus, all markers between the trait locus and the end of the chromosome have to be included.

**Arbitrary mapping functions:** On a certain haplotype, which has been observed in a fully informative half-sib, there are the hidden trait locus and, e.g., two visible flanking markers. These three loci define two intervals on the chromosome. The probability that a crossover has occurred in both intervals can be written as \( p(1, 1) \) and the probability that both intervals are free from a crossover as \( p(0, 0) \), where 0 and 1 denote no crossover and crossover in an interval, respectively. When a 1-1 haplotype has been observed and Haldane’s mapping function is applied, then \( p(1, 1) \) becomes \( \theta \) and \( p(0, 0) = (1 - \theta)(1 - \theta) \). In the formula for the probability of observing a 1-1 haplotype in an affected half-sib (\( h_i \) in Table 1), \( a_i \) can therefore be expressed as \( p(1, 1) \) and \( h_i \) as \( p(0, 0) \). An examination of Table 1 shows that all \( a_i \) and \( b_i \) values can be interpreted as Haldane crossover probabilities. The six different probabilities \( h_{\text{g}} - h_6 \) in Table 1 can therefore be rewritten by using general expressions for crossover probabilities, which include those derived from the Haldane mapping function or any other suitable mapping function as special cases (Table 3). For known map distances the probability distribution of crossover combinations in a linkage
TABLE 3

Probabilities of observing different paternal haplotypes in affected half-sibs as functions of crossover probabilities and the parameter $c$

<table>
<thead>
<tr>
<th>Paternal haplotype</th>
<th>Index $i$</th>
<th>Probability $h_i$ of paternal haplotype if individual is affected</th>
<th>$a_i$</th>
<th>$b_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1</td>
<td>1</td>
<td>$h_1 = \frac{p(0, 0) + p(1, 1) \cdot c}{1 + c}$</td>
<td>$p(1, 1)$</td>
<td>$p(0, 0)$</td>
</tr>
<tr>
<td>1-2</td>
<td>2</td>
<td>$h_2 = \frac{p(0, 1) + p(1, 0) \cdot c}{1 + c}$</td>
<td>$p(1, 0)$</td>
<td>$p(0, 1)$</td>
</tr>
<tr>
<td>2-1</td>
<td>3</td>
<td>$h_3 = \frac{p(1, 0) + p(0, 1) \cdot c}{1 + c}$</td>
<td>$p(0, 1)$</td>
<td>$p(1, 0)$</td>
</tr>
<tr>
<td>2-2</td>
<td>4</td>
<td>$h_4 = \frac{p(1, 1) + p(0, 0) \cdot c}{1 + c}$</td>
<td>$p(0, 0)$</td>
<td>$p(1, 1)$</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>$h_5 = \frac{p(0) + p(1) \cdot c}{1 + c}$</td>
<td>$p(1)$</td>
<td>$p(0)$</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>$h_6 = \frac{p(1) + p(0) \cdot c}{1 + c}$</td>
<td>$p(0)$</td>
<td>$p(1)$</td>
</tr>
</tbody>
</table>

group can be computed as described by Schnell (1961).

In fully informative half-sibs with two flanking markers there are always exactly two possible crossover combinations for a certain observed paternal marker haplotype. The first crossover combination corresponds to a gamete carrying the susceptibility allele $D$ and has probability $b_i$ and the second combination corresponds to a chromosome segment with a $d$ allele and has probability $a_i$. As already mentioned above, in half-sibs, which are not fully informative, all possible marker haplotypes at a linkage group of adjacent markers have to be considered. On a linkage group with $M$ marker loci and one trait locus between two markers there are $M$ chromosome segments between loci and $2^M$ different crossover combinations; $50\%$ of these correspond to paternal gametes with a $d$ allele and $50\%$ to paternal gametes with a $D$ allele at the trait locus. These gametes and their crossover probabilities can be enumerated and sorted in such a way that all $2^{M-1} d$-carrying gametes precede the $2^{M-1} D$-carrying gametes. After indexing the sorted gametes with $g$, the probability for observing the complementary maternal alleles at the considered linkage group can be written as

$$w_g^g = \prod_{s=1}^{M} f_{s,g},$$

where $f_{s,g}$ is defined as in the previous section. The crossover probabilities for the first group of $d$-carrying gametes (indexed from 1 to $2^{M-1}$) are denoted by $a_g$ and the crossover probabilities for the second group of $D$-carrying gametes (indexed from $2^{M-1}$ to $2^M$) by $b_g$. A value for $a_g$ can then be calculated by summing over the first $2^{M-1}$ $d$-carrying gametes,

$$a_g = \sum_{\varepsilon=1}^{2^{M-1}} w_g^g a_{\varepsilon},$$

and $b_g$ becomes

$$b_g = \sum_{\varepsilon=2^{M-1}+1}^{2^M} w_g^g b_{\varepsilon}.$$
TABLE 4

Probabilities for paternal haplotypes as functions of the transition probability $t$

<table>
<thead>
<tr>
<th>Paternal haplotype</th>
<th>Probability of paternal haplotype if individual is affected</th>
<th>$a_i$</th>
<th>$b_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1</td>
<td>$t(1-\theta_1)(1-\theta_2) + (1-t)\theta_1\theta_2$</td>
<td>$0, \theta_2$</td>
<td>$(1-\theta_1)(1-\theta_2)$</td>
</tr>
<tr>
<td>1-2</td>
<td>$\theta_2 + (1-t)\theta_1(1-\theta_2)$</td>
<td>$0, (1-\theta_2)$</td>
<td>$(1-\theta_1)\theta_2$</td>
</tr>
<tr>
<td>2-1</td>
<td>$\theta_1(1-\theta_2) + (1-t)(1-\theta_1)\theta_2$</td>
<td>$(1-\theta_1)\theta_2$</td>
<td>$\theta_2$</td>
</tr>
<tr>
<td>2-2</td>
<td>$\theta_1\theta_2 + (1-t)(1-\theta_2)(1-\theta_1)$</td>
<td>$(1-\theta_1)(1-\theta_2)$</td>
<td>$\theta_2$</td>
</tr>
<tr>
<td>1</td>
<td>$t(1-\theta) + (1-t)\theta$</td>
<td>0</td>
<td>$(1-\theta)$</td>
</tr>
<tr>
<td>2</td>
<td>$\theta + (1-t)(1-\theta)$</td>
<td>$(1-\theta)$</td>
<td>0</td>
</tr>
</tbody>
</table>

The probabilities of observing different classes of paternal haplotypes in affected half-sibs are expressed as functions of the marker-derived transition probability $t$ and the recombination rates $\theta_1$ and $\theta_2$ between the trait locus and the first and second flanking markers, respectively.

\[
1 - t = \frac{c}{1 + \exp(c)}.
\]

(48)

Therefore the different probabilities $h_i$ of observing a certain paternal haplotype with index $i$ in informative offspring, given the individual is affected and given the marker haplotypes of the sire, can be rewritten as functions of $t$ (Table 4). The general formula for the $h_i$’s is

\[
h_i = th_i + (1-\theta)a_i
\]

(49)

and the first and second derivatives of $\ln(h_i)$ are

\[
\frac{\partial \ln h_i}{\partial t} = \frac{h_i - a_i}{tb_i + (1-t)a_i}
\]

(50)

and

\[
\frac{\partial^2 \ln h_i}{\partial t^2} = \frac{-(h_i - a_i)^2}{(tb_i + (1-t)a_i)^2} = \left(\frac{\partial \ln h_i}{\partial t}\right)^2.
\]

(51)

These derivatives can be used to attain the maximum-likelihood estimate for $t$. Alternatively it is possible to transform the estimates $\hat{c} = \exp(\eta)$ into estimates $\hat{t} = 1/(1 + \hat{c})$.

The estimates of $\hat{t}$ are of interest, because the probability that two half-sibs with explanatory variables $x_1$ and $x_2$ have a probability of sharing a paternal allele identical by descent is

\[
t_{ij}t_{j2} + (1-t_{ij})(1-t_{j2}) + (t_{ij}(1-t_{j2}) + t_{j2}(1-t_{ij}) - F_{sw},
\]

(52)

which is 0.5 under the null hypothesis if the sire is not inbred ($F_{sw} = 0$). A maximum-likelihood estimation of $\hat{\beta}$ thus also provides trait- and marker-dependent estimates of all possible pairwise gametic IBD probabilities between the family members at a given map position: using $t_j = c/(1 + c) = \exp(\eta_j)/(1 + \exp(\eta_j))$ gives trait-dependent estimates and including the offset $\ln a_i - \ln b_i$, i.e., using $t_j = c_j/(1 + c_j) = \exp(\eta_j)/(1 + \exp(\eta_j))$, gives trait- and marker-dependent estimates.

**Computational considerations:** The phenomenon of unbounded parameter estimates may arise in applying the iterative equations (27) and (30). Under certain circumstances this is what we expect: if a sire is heterozygous for a recessive gene with full penetrance and $\ln c$ is estimated from affected offspring of this sire, as described above, we expect a value of either plus or minus infinity, because 100% of the affected offspring have inherited the recessive allele from the sire. When $\ln c$ tends to minus infinity, the contributions to the log-likelihood have limiting values of

\[
\lim_{\ln c \to -\infty} (\ln h_i) = \lim_{c \to -\infty} \left[\ln \left(\frac{h_i + a_i c}{1 + c}\right)\right] = \ln(h_i)
\]

(53)

and when $\ln c$ is unbounded and positive

\[
\lim_{\ln c \to +\infty} (\ln h_i) = \lim_{c \to +\infty} \left[\ln \left(\frac{h_i + a_i c}{1 + c}\right)\right] = \ln(a_i).
\]

(54)

This implies that for practical purposes unbounded parameter estimates can be handled by restricting the elements of $\hat{\beta}$ to a range of $-20$ to $+20$ and using a maximum value of 20 for $|\eta|$ This should yield values for the log-likelihood that are close to the limit and sufficiently precise for purposes of hypothesis testing.

**EXAMPLE**

An example demonstrates the application to so-called super sisters, as they occur in the European honeybee *Apis mellifera*. In honeybees some traits, e.g., honey yield, can be measured at colony level only. For other traits, which can be measured on individual worker bees, it may be possible to collect a rather large number of individuals showing a certain phenotype. This becomes possible even for a rare phenotype, because a colony comprises tens of thousands of offspring from a single queen. Such traits could be morphological or, e.g., defense behavior against the parasitic mite *Varroa jacobsoni*. If, in an experiment, the queen has been artificially inseminated with the sperm of a single haploid drone (Harbo 1989), then all workers share the same paternal
TABLE 5

Super sister example data

<table>
<thead>
<tr>
<th>Genotypes</th>
<th>Observations</th>
<th>Maternal haplotype</th>
<th>Index 1</th>
<th>Index 2</th>
<th>Probability 1</th>
<th>Probability 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1, 2 1, 3 1, 1</td>
<td>6</td>
<td>1-1-1</td>
<td>1</td>
<td>1</td>
<td>0.3253</td>
<td>0.1335</td>
</tr>
<tr>
<td>1, 2 1, 3 1, 2</td>
<td>7</td>
<td>1-1-2</td>
<td>1</td>
<td>2</td>
<td>0.1747</td>
<td>0.1295</td>
</tr>
<tr>
<td>1, 2 2, 3 1, 1</td>
<td>3</td>
<td>1-2-1</td>
<td>2</td>
<td>3</td>
<td>0.1747</td>
<td>0.2199</td>
</tr>
<tr>
<td>1, 2 2, 3 1, 2</td>
<td>8</td>
<td>1-2-2</td>
<td>2</td>
<td>4</td>
<td>0.3253</td>
<td>0.5171</td>
</tr>
<tr>
<td>2, 2 1, 3 1, 1</td>
<td>3</td>
<td>2-1-1</td>
<td>3</td>
<td>1</td>
<td>0.3253</td>
<td>0.1335</td>
</tr>
<tr>
<td>2, 2 1, 3 1, 2</td>
<td>2</td>
<td>2-1-2</td>
<td>3</td>
<td>2</td>
<td>0.1747</td>
<td>0.1295</td>
</tr>
<tr>
<td>2, 2 2, 3 1, 2</td>
<td>12</td>
<td>2-2-1</td>
<td>4</td>
<td>3</td>
<td>0.1747</td>
<td>0.2199</td>
</tr>
<tr>
<td>2, 2 2, 3 1, 2</td>
<td>27</td>
<td>2-2-2</td>
<td>4</td>
<td>4</td>
<td>0.3253</td>
<td>0.5171</td>
</tr>
</tbody>
</table>

Sixty-eight workers are descending from a single queen and a single drone. The queen has genotype 1, 2 at all three marker loci and the haploid drone has marker alleles 2, 3, and 1. The genotypes of the worker bees at the first, second, and third marker are given in the first column, the number of workers with the same genotype in the second column, and the maternal haplotypes in the third column. For all workers index 1 and index 2 indicate the class of maternal haplotype for the first and second marker brackets, respectively. Probabilities 1 and 2 denote the contribution to the likelihood of a single individual computed for the map position at 52 Haldane cM and \( c \) values of 1 and 5.234, respectively.

In a hypothetical example tissue from 68 super sisters with the same rare phenotype is available together with tissue of the queen and the drone. If all these animals are genotyped for codominant markers, then the super sisters can be treated like fully informative half-sibs in the analysis because the haploid drone always transmits the same allele. The marker genotypes of the worker bees are given in Table 5. The queen is heterozygous 1, 2 at all markers. The drone’s genotypes at the three consecutive marker loci are 1, 3, and 2, respectively, and allow us to resolve the genotypes of the worker bees unequivocally into maternal haplotypes (Table 5). The frequencies of these maternal haplotypes allow us to assume that the haplotypes of the queen are 1-1-1 and 2-2-2. The number of super sisters with a particular maternal haplotype is given in the second column of Table 5. The fourth and fifth columns of this table show the index (as defined in Table 1) for the class of the maternal haplotype for the first and the second marker bracket separately. The contributions of single individuals to the likelihood \( h_i \) are given for the null hypothesis (\( \epsilon = 1 \)) and also for the maximum-likelihood estimate \( \epsilon = 5.234 \) at a map position of 52 cM. The length of the first marker interval was assumed to be 35 Haldane cM and for the second one 60 Haldane cM. The LOD-score profile over both marker brackets is given in Figure 2. It shows that the most probable locus position is at 52 cM.

The maximum-likelihood estimate of \( \epsilon \) translates to an identity-of-descent probability of 0.73 for the maternal gamete at 52 cM, when the marker genotypes are unknown for two randomly chosen affected individuals. Table 6 shows the same IBD probability for all possible pairs of affected progeny with known flanking markers, with values ranging from 0.17 to 0.97.

Under natural conditions a queen mates with several drones. Each of them is the founder of a group of super sisters and therefore the entire colony is a mixture of several super sister families with a common mother. Tissue from the drones is usually not available, because the matings take place far away in the air. Under these circumstances a trait-based linkage analysis with worker bees can be conducted in the same way as for half-sibs, with some of the workers only partially informative. The application of artificial insemination, however, enables us to save tissue from all the drones, regardless of their number. With artificial insemination and semen from more than a single drone, markers can be used for a paternity test to determine the membership of each genotyped worker bee to one of the super sister groups in the colony. Then, by genotyping all drones together with the queen and the workers, the resolution of the progeny’s marker genotypes into maternal haplotypes again becomes unequivocal and the analysis can proceed as described in the example for a single super sister family.

DISCUSSION

In applying the method to partial- or whole-genome scans it becomes necessary to account for repeated testing. The simplest method would be to use a predefined significance threshold, e.g., a LOD score of 3 with 1 d.f. (Ott 1991). In the important case of genotyping affected family members only, a permutation test (Churchill and Doerge 1994) cannot be applied, because all progeny remain to have the same phenotype before and after permutation. The distribution of the
test statistics under the null hypothesis of no linkage can, however, be achieved by simulation. For this purpose parental marker alleles for each progeny can be drawn from the known haplotypes of the parent of each half-sib family by a random walk along the chromosome and gene drop. The missing alleles of the other parent can be supplemented by choosing them randomly from the population, according to their gene frequency. Of course it is not necessary to perform the last step of this simulation in fully informative half-sibs (e.g., with a cross of inbred lines) or super sisters. If all members of a half-sib family are genotyped, regardless of their phenotype, then a permutation test can be applied. The same is true with selective genotyping, provided that all possible paternal haplotypes are adequately represented and a \( c \) value of 1 can be expected in the joint data and, on average, also in the permuted data. This would be the case, for example, in a daughter design where marker genotypes are determined for the best 25% and the worst 25% of the animals.

The null hypothesis \( H_0: c = 1 \) will usually be adequate in all analyses with only a single phenotypic class in one family. If marker data are available for two distinct phenotypes, then the null hypothesis could either be \( H_0: c = c = 1 \) or \( H_0: c = c = 0 \). The first hypothesis would also be rejected in the case of segregation distortion. With a twofold excess of the frequency of the second paternal haplotype in both phenotypes we would, e.g., expect \( c_1 \) and \( c_2 \) to be equal with a value of 2. Therefore a test for linkage should be performed as a test for differences in the \( c \) values of different phenotypes or, in other words, as a test for the trait dependency of \( c \). It may, however, be possible, at least in principle, to compare the two alternative null hypotheses by a likelihood-ratio test, to draw inferences on the presence or absence of segregation distortion.

The linear predictor \( \eta = \chi'\beta \) adds a lot of flexibility to the analysis: e.g., with multiple phenotypic categories \( \beta \) contains a separate parameter for each of those categories or a linear regression coefficient if the phenotype is continuous. A \( \beta \) with interactions between families would be appropriate for a situation with several unrelated half-sib families, where some of the sires may be heterozygous and others not and the linkage phase between markers and the susceptibility locus is unknown \textit{a priori}. The sign of \( \eta \) indicates which paternal haplotype carries the susceptibility locus: if \( \eta \) for affected animals is greater than zero then the corresponding \( c \) is larger than one and the second haplotype of the sire carries the \( D \) allele; otherwise the reverse is true. The linear predictor also allows for the inclusion of several traits, no matter if these traits are categorical, continuous, or a mixture of both. Another interesting case is a trait

**TABLE 6**

Pairwise IBD probabilities at 52 cM on the maternal gametes for two randomly chosen affected individuals with known maternal marker haplotype

<table>
<thead>
<tr>
<th>Maternal marker haplotype of second individual</th>
<th>Maternal marker haplotype of first individual</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1-1</td>
</tr>
<tr>
<td>1-1</td>
<td>0.73</td>
</tr>
<tr>
<td>1-2</td>
<td>0.37</td>
</tr>
<tr>
<td>2-1</td>
<td>0.22</td>
</tr>
<tr>
<td>2-2</td>
<td>0.17</td>
</tr>
</tbody>
</table>

Recombination rates between the trait locus and the first and second markers are 0.150 and 0.285, respectively.
with observed values in two classes, where in the first class all animals have the same trait value without any variability (e.g., 100% of the coat is white in albinos) and in the second class the trait is variable (e.g., the percentage of white coat is <100 and may reach 0 in non-albinos). An adequate linear predictor would comprise three parameters: a mean for each of the two classes plus a regression on the trait value only for animals of the second class.

The proposed maximum-likelihood approach may be viewed as a generalization of the logistic regression procedure of Henshall and Goddard (1999). Under perfect linkage and in fully informative half-sibs the numbers of animals with a paternally derived Q allele and a paternally derived q allele at a QTL can be directly observed. Henshall and Goddard (1999) showed that the trait dependency of these counts can be exploited in a logistic regression analysis. For recombination rates larger than zero the authors suggested that the regression coefficients obtained at the closest markers be interpolated and that expected counts of Q and q animals (derived from the markers) be used to calculate the likelihood-ratio test. It has been demonstrated in this article that these approximations are not necessary when paternal marker haplotypes and their frequencies are used as observations instead of the, in most cases, unobservable number of Q and q individuals in a family. The range of possible applications and the biological interpretation of the estimates obtained by logistic regression and the maximum-likelihood approach of this article are, of course, identical. The analysis is “nonparametric” in the sense that there are no parameters, which explicitly describe the mode of inheritance, such as penetrances and gene frequencies. It is, however, possible to derive the c value if the mode of inheritance can be specified in detail. As already mentioned above, the c value for affected animals tends toward infinity for a single recessive gene (or zero, depending on the linkage phase). This holds for any value of the penetrance of the homozygous recessive genotype. The use of the limiting values of the contributions to the log-likelihood given in (53) and (54) makes it possible to test for the recessive expression of the trait locus by comparing the likelihood of the data for an infinitely large c value with the maximum of the likelihood, which has been estimated from the data. A significant result would exclude all kinds of recessive single-gene inheritance, regardless of the penetrance of the homozygous recessive genotype. Another interesting pattern is a single, fully dominant gene. In this case the c value for affected animals becomes equal to the gene frequency (or one over the gene frequency for the alternative linkage phase).

Observing the expected c value for different phenotypes is also useful for power calculations and helps in answering the question of which phenotypes should be genotyped preferentially. Usually the c value for a group of unaffected animals will be larger than the c value for a group of unaffected animals and the experimental power from genotyping a certain number of affected will be higher than from genotyping either the same number of unaffected or a mixture of affected and unaffected. After having obtained a significant linkage, the genotypes of unaffected half-sibs are, however, useful for excluding the existence of segregation distortion, which could lead to false positive results, if only affected sibs are genotyped. In the special case when a sire is heterozygous for a dominant trait locus and is mated to dams from a line in which the recessive allele is fixed, then the c value for the first chromosome and affected progeny becomes equal to the c value for the second chromosome and unaffected animals, and animals of both phenotypes provide the same possibility to map the gene. There may be further examples for such an equivalence, especially in cases with more than a single trait locus.

In many species, e.g., cattle, pigs, honey bees, trees, and fish, the reproductive capacity of one or both sexes is tremendously higher compared to humans, either naturally or by the extensive use of artificial insemination in commercial or experimental populations. In human genetics methods for mapping disease loci by genotyping affected sibs for multiple markers have been implemented, e.g., in the widely used “Genehunter” program (Kruglyak et al. 1996) or derivatives such as “Allegro” (Gudbjartsson and Jonasson 1999). The author’s experience is that these implementations are at present technically restricted to half-sib families sized not larger than ~20. The proposed trait-based maximum-likelihood analysis makes it possible to analyze simultaneously collections of large and even very large half-sib families comprising only affected individuals practically without any important restrictions due to family size. Beyond this important class of mapping experiments, the range of potential applications is much broader and offers possibilities for linkage detection with any multilocus feasible mapping function in experiments with features such as selective or nonselective genotyping, single or multiple traits, continuous or discontinuous traits, and combinations of them.

LITERATURE CITED


Kruglyak, L., J. D. Mark, M. P. Reeve-Daly and E. Lander, 1996

Communicating editor: C. Haley