Trio Logic Regression - Detection of SNP-SNP Interactions in Case-Parent Trios

Qing Li[1], Thomas A Louis[1], M Daniele Fallin[2], Ingo Ruczinski[1,*]

Deptartment of Biostatistics[1] and Epidemiology[2]

Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21218.

June 8, 2009

* Author to whom correspondence should be addressed (ingo@jhu.edu)

Abstract

Statistical approaches to evaluate higher order SNP-SNP and SNP-environment interactions are critical in genetic association studies, as susceptibility to complex disease is likely to be related to the interaction of multiple SNPs and environmental factors. Logic regression (Kooperberg et al., 2001; Ruczinski et al., 2003) is one such approach, where interactions between SNPs and environmental variables are assessed in a regression framework, and interactions become part of the model search space. In this manuscript we extend the logic regression methodology, originally developed for cohort and case-control studies, for studies of trios with affected probands. Trio logic regression accounts for the linkage disequilibrium (LD) structure in the genotype data, and accommodates missing genotypes via haplotype-based imputation. We also derive an efficient algorithm to simulate case-parent trios where genetic risk is determined via epistatic interactions.

Keywords: Case-parent trios, interaction, logic regression, single nucleotide polymorphisms.

1 Introduction

Statistical approaches to evaluate higher order interactions between SNPs, or between SNPs and environmental variables, are critical for analyzing complex diseases as higher susceptibility is likely to be related to the interaction of multiple SNPs and environmental factors. The effect sizes seen in complex diseases are typically very small, and therefore the power to detect those small effect sizes can crucially depend on whether methods to simultaneously investigate SNPs and environmental variables are employed, i.e. whether or not such interactions are directly assessed. This, however, creates statistical challenges in the analysis of how SNPs and environmental variables relate to the disease outcome, since the number of possible interactions between genetic markers and the environmental factors is immense. To address this issue, many tools from the statistical and machine learning literature, developed to deal with high-dimensional search spaces, have been applied to multi-marker SNP data, for example neural networks (Lucek and Ott, 1997; Bhat et al., 1999; Ritchie et al., 2003b; North et al., 2003; Tomita et al., 2004), random forests (Breiman, 2001; Lunetta et al., 2004; Bureau et al., 2005; Chen et al., 2007), and various other methods based on partitions, trees, and splines, and ensembles of base learners (e.g. Chen et al., 2003; Cook et al., 2004; Zhang et al., 2008). Some approaches to delineate higher order interactions were specifically developed for SNP data, such as the multifactor dimensionality reduction techniques (Hahn et al., 2003; Ritchie et al., 2003a; Moore, 2004; Ritchie and Motsinger, 2005; Ritchie, 2005), the restricted partition method (Culverhouse et al., 2004, 2007), and logic regression (Kooperberg et al., 2001; Ruczinski et al., 2003, 2004). An overview and comparisons of some of these algorithms can for example be found in McKinney et al. (2006), Heidema et al. (2006), and Vermeulen et al. (2007), and additionally, an extensive discussion of the properties of these algorithms is given in Musani et al. (2007).

In logic regression, the interactions between SNPs and environmental variables are assessed in a regression framework, where interactions become part of the model search space. Given a set of binary covariates, logic regression creates new predictors for the response by considering

Boolean combinations of such binary covariates, while also being able to adjust for other covariates of interest as main effects in the regression model. The SNPs are usually recorded as two indicator variables under dominant or recessive coding, which enables a statistical representation of many genetic models (for example, double penetrance models) and biological interactions of interest. The regression framework allows for quantitative statements such as *the odds of disease for a subject who has at least one variant allele at both SNP 7 and SNP 12 are three times higher compared to a subject of the same age who does not have variant alleles at both of these loci.* The model search is carried out using a simulated annealing algorithm, and model selection is performed via cross-validation and permutation tests. Logic regression has been applied in numerous SNP association studies, for example on breast cancer (Justenhoven et al., 2008), colorectal cancer (Suehiro et al., 2008), prostate cancer (Etzioni et al., 2004), bladder cancer (Andrew et al., 2008), head and neck squamous-cell carcinoma (Harth et al., 2008), hypertension (Huang et al., 2004), and myocardial infarction and ischemic stroke (Enquobahrie et al., 2008). Logic regression has also been applied in the context of other biomedical research, such as the detection of transcription factor binding sites (Keles et al., 2004), DNA methylation (Feng et al., 2005), HIV replication capacity (Segal et al., 2004), immunohistochemistry (Yaziji et al., 2006), and biomarker detection (Vaidya et al., 2008). Extensions of the logic regression methodology exist to detect associations between disease and haplotypes in blocks with little or no recombination (Clark et al., 2007), to generate ensembles of plausible models based on Markov chain monte carlo algorithms (Kooperberg and Ruczinski, 2005), and to obtain measures of variable importance and approaches suitable for variable selection (Schwender and Ickstadt, 2008).

The above cited methods, including logic regression, are typically employed in case-control or cohort studies, relating genotypes to binary, ordinal, or numeric phenotypes. In addition to population based designs, family based studies offer an appealing alternative, since these designs are robust against population substructure, and allow for the assessment of linkage and association (Spielman and Ewens, 1996; Gauderman et al., 1999; Fallin et al., 2002; Laird and Lange, 2006).

Arguably, the simplest and most prominent test is the transmission disequilibrium test (TDT), a completely non-parametric approach applicable to trios of parents plus affected offspring (Spielman et al., 1993). Many extensions to the TDT have been proposed, for example allowing for multi-allelic and haplotype based tests, general pedigrees, missing data, quantitative traits, many using parametric approaches (see for example Laird and Lange, 2008, for a comprehensive review).

Much thought and effort was provided by the community to devise methods to detect gene-gene and gene-environment interactions in family data, particularly for nuclear families, and most prominently among those, for parentaffected-child trios (e. g. Schaid, 1999; Lunetta et al., 2000; Culverhouse et al., 2002; Lanktree et al., 2004; Baksh et al., 2006, 2007; Kotti et al., 2007a,b). However, these approaches do not provide algorithms to allow for the actual search of SNP-SNP interactions, but quantify the statistical significance of candidate interactions under investigation. A noticeable exception to this is the algorithm MDR-PDT (Martin et al., 2006). MDR-PDT allows for the search and assessment of higher-order interactions by merging the genotype-pedigree disequilibrium test (Martin et al., 2003) with the above mentioned multifactor dimensionality reduction technique (Ritchie et al., 2003a): multilocus genotypes are pooled into high-risk and low-risk groups, sequentially reducing the dimensionality of the predictors.

In this manuscript, we extend the logic regression methodology, employing probabilistic search algorithms (simulated annealing) to detect and assess higher order interactions in trios with affected probands. Trio logic regression also accounts for the linkage disequilibrium (LD) structure in the genotype data, and accommodates missing genotypes via haplotype-based imputation. We also derive an efficient algorithm to simulate case-parent trios where the genetic risk is determined via epistatic interactions.

2 Methods

Trio logic regression

Logic regression is an adaptive regression and classification tool introduced by Ruczinski et al. (2003) to address problems arising when data of mostly binary covariates are analyzed, and the interactions between those predictors are of main interest. Given a set of binary covariates, logic regression creates new predictors for the response by considering Boolean combinations of the binary covariates. Logic regression models are of the form

$$g(E[Y]) = \alpha + \sum_{i=1}^{p} \gamma_i Z_i + \sum_{j=1}^{t} \beta_j L_j \tag{1}$$

where g is an appropriate link function for the response Y, $Z_1, \ldots, Z_k$ are covariates included as additive terms (this set can be empty), and $L_1, \ldots, L_t$ are Boolean terms of the binary covariates $X_1, \ldots, X_k$ such as $L_j = (X_2 \vee X_4) \wedge X_7$. In genetic association studies, the SNPs are usually recorded as two indicator variables under dominant or recessive coding, which enables a statistical representation of many genetic models (for example, double penetrance models) and biological interactions of interest. This framework allows then for statements such as *comparing two smokers of the same age, the odds of disease for the person with two variant alleles at SNP7 and at least one variant alleles at SNP44 are three times higher than for the other person without this genetic pattern.* The logic regression framework includes many forms of regression (such as linear and logistic regression, Cox proportional hazards model, and more). In general, any type of model can be considered, as long as a scoring function (such as a deviance or likelihood) can be defined. The model search is carried out using a simulated annealing algorithm, and model selection is performed via cross-validation and permutation tests. A detailed description is given in Ruczinski et al. (2003).

We extend the logic regression methodology to accommodate the case-parent design by setting up the trio data for conditional logistic regression with a 1:3 matching ratio for the case genotypes versus 3 possible Mendelian realizations given the parents. At each locus, one of four possible pairs of parental alleles is transmitted to the affected offspring. The other (unobserved) possible Mendelian realizations of genotypes for a child, given the parental genotypes, can be used as artificial controls (the "pseudo-controls"), and the resulting matching structure can be accounted for using a conditional likelihood in a marker-by-marker analysis (Self et al., 1991; Schaid, 1996). This approach is known as the genotypic transmission disequilibrium test (TDT). Logic regression however needs to consider all markers simultaneously to detect higher order SNP-SNP interactions. This is problematic since the number of possible pseudo-control genotype sets in particular grows exponentially with the number of unlinked markers considered since all locus combinations are possible (e. g., for n unlinked markers, 3^n pseudo-controls are possible). We avoid this dimensionality problem by restricting the analyses to 1:3 matching. For unlinked markers, we choose a random order for the three possible pseudo-controls at each marker, and concatenate these genotypes. For markers in tight LD (according to some block definition using measures such as $D^{'} or R^2$), we first generate pseudo-controls for the entire haplotype block, and then sample (without replacement) three realizations of possible pseudo-controls from this entity. We will discuss the details how to accommodate missing genotypes in a later section (page 12 ff), and summarize the steps how to generate a data frame suitable for trio logic regression:

1. Estimate the haplotype blocks and the haplotype frequencies using the parents' genotypes.
2. For each block and each trio, sample haplotype pairs for the parents and the offspring consistent with the observed genotypes in the trio, allowing for missing data.
3. Generate the probands genotype data from the haplotypes that were passed from the parents.
4. For each block and each trio, generate genotypes for three pseudo-controls (PC1, PC2, PC3) using the parents' haplotypes that were not passed to the proband. The assignment to PC1, PC2, and PC3 is random.

5. Assemble three pseudo-controls for each trio by augmenting the genotypes from the blocks.
6. For each locus, translate the genotype data into two binary variables in dominant and recessive coding.

As any sampling based imputation procedure, the approach outlined above generates one complete set of data, which can for example be used in logic regression to search for higher order SNP-SNP interactions. However, when formal inference is the objective, and estimating the effects size and its standard error is of primary interest, the extra level of uncertainty due to missing data should be acknowledged. This can be done, for example, by means of multiple imputation (Rubin, 1996; Schafer, 1999). Several complete data sets are constructed, and the final inference is based on the parameter estimates and standard errors derived from the individual completed data sets, taking the variability of the estimates within a data set and between all data sets into account.

Logic regression embedded in a conditional logistic regression framework can then be run to find the best scoring models (e. g. , the models with the lowest deviances) for models of different sizes. In this manuscript, we limit ourselves to models with only one Boolean term, and the model size is defined as the number of predictors in the Boolean term. The parameters in the models are estimated simultaneously in the model search, however, some special attention has to be given to possible non-convergence in the optimization procedure. This issue can best be seen when considering the conditional logistic regression likelihood, which can be written as (Breslow et al., 1978)

$$L(\beta) = \prod_{i=1}^{N} \left(\frac{\exp(X_{i0}\beta)}{\exp(X_{i0}\beta) + \Sigma_{m=1}^{3} \exp(X_{im}\beta)} \right) \tag{2}$$

In this setting, i refers to a trio ($i \in \{1, \ldots, N\}$), X_{i0} refers to the exposure of the proband in trio i, and (X_{i1}, X_{i2}, X_{i3}) are the exposures of the 3 pseudo-controls in trio i. In this setting, the likelihood will not have a maximum if and only if for either all the trios where $X_{i0} = 0$ or all the trios where $X_{i0} = 1$, the respective pseudo-controls exposures are equal to the probands' exposures, i. e. $X_{i1} = X_{i2} = X_{i3} = 0$ for all trios with $X_{i0} = 0$, or $X_{i1} = X_{i2} = X_{i3} = 1$ for all trios with $X_{i0} = 1$ (see Appendix). In other words, for either the exposure or the non-exposure groups, all pseudo-controls are equal to the respective probands. This might occur in a particular sample where the number of trios is small and the (or some of the) SNPs contributing to the exposure definition have extremely low minor allele frequency, however, it is not plausible that such a separation would exist in the population in truth. Thus, it is necessary from a computational perspective (to avoid non-convergence) and meaningful from a scientific perspective to exclude these settings in the evaluation of logic regression models. We thus implemented a check that rejects any such models in the annealing algorithm.

Special attention needs also to be given to model selection to avoid over-fitting. In logic regression, some definition of model size is required, and typically the total number of predictors in the Boolean term is used as such. However, models of different sizes are not nested. In particular, models with equal numbers of Boolean terms employ the same number of parameters, and thus, common measures of model complexity do not apply. This was recognized in the original methodology, and special model selection techniques such as sequential permutation tests devised (Ruczinski et al., 2003). However, these are also not applicable in trio logic regression due to the conditional likelihood. Therefore, a modified permutation test is proposed.

Similar to the original methodology, sequential hypothesis tests are carried out, and the score (such as the likelihood) of a larger model using the original data is compared to scores derived using the permuted data, conditioning on models of various sizes (see Ruczinski et al., 2003, for details). However, the conditioning on the best model of a certain size now has to take the grouping of the case / pseudo-control quartet into account. In particular, when conditioning on any particular model, the likelihoods of the original and the permuted data have to be identical. This can be achieved by swapping the case with one of its pseudo controls that has the same exposure status. From Table 12 it follows immediately that the likelihood does not change under this procedure. For d_0 and d_1 we have three possibilities for each proband to select a pseudo-controls for swapping, for c_0 and c_1 we have two possibilities, for b_0 and b_1 there is only one possibility. Trios in which the proband has a different exposure than all of its pseudo-controls are not altered. As in the original methodology, the best score of a larger model is compared to a sequence of permutation distributions, conditioning on models of increasing sizes. A shift in those permutation distributions towards higher likelihoods indicates signal in the data. Overfitting starts when the permutation distributions stop shifting and the score of the larger model derived using the original data resembles those in the permutation distribution.

Case-parent trio simulation

We next describe an efficient approach to simulate case-parent trios, which is a key ingredient to validate the logic regression methodology and software. However, simulations based on haplotypes and mating tables very often result in computationally intractable problems. We show that a naive enumeration of mating patterns quickly results in unworkable dimensionsonalities, and introduce an alternative that makes the required computations feasible. Using the here introduced notation also simplifies the description for the method trio logic regression handles missing genotypes, which will be discussed in the following section. The following simulation methods have been implemented in the function `trio.sim()` in the R package `trio`.

We assume two risk groups in the population defined by some genotype pattern G, and assume that the probability of disease p is given via the log-odds as

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta I_G, \tag{3}$$

where I is the indicator function, and α and β are some fixed parameters. The genotype patterns here are based on Boolean combinations of SNPs in dominant and recessive coding, such as ($\text{SNP}_1^R \wedge \text{SNP}_{15}^D$), indicating that subjects with two variant alleles at SNP 1 and at least one variant allele at SNP 15 are at higher risk (assuming $\beta > 0$).

To simplify notation, we follow Weinberg et al. (1998) and use the letters F, M, and C to represent the haplotype pairs (diplotypes) of the father, the mother and the child. We refer to the joint probability distribution of F, M and C as the mating table. Further, we use the letter D to indicate an affected proband. To simulate case-parent trios, we therefore need to specify $P(F, M, C|D)$ for each haplotype block (assuming independence between blocks). For haplotype blocks that do not contain a locus involved in the genetic signal that defines the high risk group in equation (3), the haplotype frequencies in the trios do not depend on the disease status of the child. Thus, we have $P(F, M, C|D) = P(F, M, C)$ for a haplotype block that does not contain information about the disease risk. Further, $P(F, M, C) = P(M, F) \times P(C|M, F) = P(M) \times P(F) \times P(C|M, F)$, under the assumption of random mating. Thus, for any block that does not contain information about the disease risk, we can sample the genotypes for the trios by randomly selecting two haplotypes for each parent using the population frequencies, and generate the proband's diplotype from the parents' haplotype pairs assuming independent segregation.

To generate the trio genotypes for the blocks that do contain information about disease risk, we have to take into account that the haplotype frequencies in those blocks are different from the population at large if we condition on having an affected proband. However, enumerating the entire mating table and calculating all these probabilities is prohibitive. Assume for example that there are three blocks with loci that contribute to the disease risk, with five haplotypes each. Thus, there are $5^3 = 125$ possible haplotype sets for a subject, resulting in $125 \times (125 + 1)/2 = 7875$ possible diplotypes for these three blocks combined. For a pair of parents we the have $7875 \times (7875 + 1)/2 \approx 31$ million haplotype pair combinations. Thus, we pursue a different strategy that avoids generating the whole mating table. We first sample the haplotype pairs for affected probands, and then sample the parents' haplotype pairs given the proband's diplotype. In particular, we use

$$P(F, M, C|D) = P(F, M|C, D) \times P(C|D) = P(F, M|C) \times P(C|D) \tag{4}$$

That is, we first sample a diplotype for an affected proband, and then sample the parents given the affected proband. We can avoid extensive enumerations by taking advantage of the fact that there are only two risk groups. The procedure is best explained in an example.

Assume that there are two haplotype blocks, with two and three loci respectively, that specify the disease risk (Table 1) as follows:

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta \times I\left[(\mathsf{SNP}_1^R \wedge \overline{\mathsf{SNP}_3^D}) \vee \mathsf{SNP}_5^R\right] \tag{5}$$

[TABLE 1 ABOUT HERE]

Thus, subjects with two variant alleles at SNP number five (the third locus in block 2) are at higher risk, and subjects with either two variant alleles at SNP number one (the first locus in block 1), or no variant alleles at SNP 3 (the first locus in block 2). We assume 3 possible haplotypes in block 1 and 4 possible haplotypes in block 3, resulting in 12 possibilities for a haplotype spanning both blocks (Table 2). Therefore, there are $12 \times (12+1)/2 = 78$ possible diplotypes in our population. Out of those 78 diplotypes, 22 have the risk genotype combination as specified above (Table 3), and we tabulate the haplotype pairs accordingly in high and low-risk diplotypes (Table 4). For computational efficiency and convenience, we also differentiate between identical and non-identical haplotypes in a diplotype (denoted as strata HR_1, HR_2, LR_1, and LR_2), and further stratify by the possible indices of the first haplotype we will sample in a pair. Note that this step needs to be carried out only once, when the simulation is initialized.

[TABLE 2 ABOUT HERE]

[TABLE 3 ABOUT HERE]

[TABLE 4 ABOUT HERE]

To reflect the diplotype distribution among affected probands, we have to consider the probability of disease given the genotype, and we note that $P(C|D) \propto P(D|C) \times P(C)$. Moreover, for any two subjects within the same diplotype stratum as defined in Table 4, we have $P(D|C_1) = P(D|C_2)$. Thus, if S denotes the union of all diplotypes within a stratum, we have

$$P(S|D) \propto P_S(D) \times P(S), \tag{6}$$

where $P_S(D)$ is the probability of disease, equal for every element in S. Thus, we can use the set of probabilities $P(S|D)$ as the sampling probabilities to pick a stratum in initial step in the sampling procedure. We then sample a haplotype (index i) using the strata probabilities as indicated in Table 4, and then pick a second haplotype (index j) from the set of permissible haplotypes, using the haplotype frequencies from this set.

With the diplotype of the affected proband available, the joint diplotype distribution for the parents $P(F, M|C)$ is rather straightforward. For computational efficiency, we again divide the joint parental distributions into strata (Table 5). We distinguish the cases when the proband's haplotype pairs are identical or not. The strata probabilities are then used to sample the actual haplotype pairs for the parents (Table 6). Two considerations are important in the calculation of the sampling probabilities. Since there is no ordering among the parents, the joint probability for a non-identical pair of diplotypes is multiplied by the factor 2. Further, depending on the parents' haplotype combinations, the number of distinct diplotypes among the four "Mendelian" children can be one, two, or four. To calculate the correct sampling probabilities, these issues are taking into consideration by multiplying the parents' diplotype probabilities with the respective factors (Table 6). In summary, given the child's diplotype, we will first sample the stratum (Table 5), and then sample the pairs for the parents within each stratum (Table 6).

[TABLE 5 ABOUT HERE]

[TABLE 6 ABOUT HERE]

Missing genotype imputation

Logic regression and trio logic regression require complete data, and thus, an imputation method for missing genotype data is needed that takes the block structure and the phase information into account. The completed trios will then be used to generate the pseudo-controls, resulting in a complete data set suitable for trio logic regression. Our imputation method is based on haplotypes and the observed trio genotype data, and we assume that these data are free of Mendelian errors (each set of three genotypes that gives rise to a Mendelian error can for example be replaced by missing values and is an option in the function `trio.check()` in the R package `trio`, though more efficient approaches might be applicable). As before, the enumeration of all possible haplotype combinations for the trios can be prohibitive, and we employ some of the previously introduced techniques for trio simulation.

We distinguish six different scenarios for missing genotypes in the trios (Table 7). In general, the genotypes of one or more subjects in the trio might be completely missing, for example if the father of the proband was unknown or not genotyped, or if the extracted DNA from one of the subjects was compromised. Otherwise, the genotypes are typically observed completely, or with low missingness rates due to genotyping problems. When the proband's genotypes are completely missing (scenarios 1 to 3), there are no restriction on the parental diplotypes other than through the observed genotypes. In these settings, we can directly sample the parental haplotypes given the observed genotype data, and derive the proband's haplotypes. Likewise, if all three family members were genotyped and have complete or partial genotype data (scenario 6), the same approach remains viable. We find the possible haplotypes for both parents given the observed data, and derive the proband's haplotypes for each of these possibilities, subject to the condition that the proband's diplotype is in agreement with the observed genotypes.

[TABLE 7 ABOUT HERE]

Computational difficulties arise when the proband's genotypes are (partially) observed, but one or both parental genotypes are missing completely (scenarios 4 and 5). In this setting all possible haplotypes need to be considered, and a technique based on the previously introduced conditional mating table circumvents the computational difficulties (Table 7). If both parental genotypes are missing completely (scenario 4), we use the following equation:

$$P(F, M, C|C_G) = P(F, M|C)P(C|C_G) \tag{7}$$

Here, the subscript G refers to "genotype", to distinguish the observed genotypes from the haplotypes. The sampling probability calculations for $P(F, M|C)$ are similar to the ones introduced in our simulation method, except that the possibilities of the child's haplotypes are now constrained by the observed genotypes. Therefore, we only need to sample the proband's haplotypes using the standardized diplotype frequencies. Once the diplotype C for the proband is fixed, the parental haplotypes are sampled exactly the same way as in the simulation approach.

The imputations are slightly more complicated if the genotype data are missing completely for one parent only. In this instance (without loss of generality, assume the father has no observed genotypes), the child's haplotype frequencies further need to be adjusted according to the mother's observed genotypes and diplotype possibilities. Unfortunately, a similar approach as described in equation (7), to first assess the diplotypes for the subjects with observed genotypes, does not work here: the joint diplotype probabilities for the mother and proband $P(M, C|M_G, C_G)$ do depend on the father's diplotypes. Thus we need to calculate the diplotype probabilities by considering the parental diplotypes that are in agreement with the observed parental genotypes, and for which at least one Mendelian offspring is in agreement with the observed proband genotypes. However, these joint probabilities can efficiently be computed. When the father's genotype is completely missing, we consider the trio's diplotype possibilities and stratify them by the mother/child combinations, grouping identical and homogeneous haplotypes pairs (Table 8). For each mother/child diplotype combination that is not in conflict with the observed genotypes, we list the corresponding scenarios and sampling weights, and calculate the sampling probabilities for all relevant diplotype combinations, and the mother/child diplotype strata. To impute the data, we sample the mother/child diplotype pair first, and then sample the paternal diplotype given the mother/proband haplotype pairs.

[TABLE 8 ABOUT HERE]

These imputation methods have been implemented in the function `trio.imp()` in the R package `trio`. We demonstrate the underlying algorithm with a very simple example using a LD block of length 3 with eight possible haplotypes (Table 9). We note though that the gain in computational efficiency increases with haplotype block length, as the number of possible haplotypes is 2^n for a block of length n. Assume that in our example the observed genotypes are 22/NA/22 for the proband, and NA/22/NA for the mother. Thus, given the observed genotype, the offspring's diplotype must only include haplotypes 6 and 8. For the same reason, the mother's diplotype must only include haplotypes 3, 4, 7, and 8. Since one maternal allele gets passed to the offspring, this can only be the allele 8. Thus the only possible diplotypes for the child are (6,8) and (8,8), and the only possible diplotypes for the mother are (3,8), (4,8), (7,8), and (8,8). If the offspring is (6,8), the father's diplotype must include haplotype 6. If the offspring is (8,8), the father's diplotype must include haplotype 8. The possibilities are easily enumerated and the respected sampling probabilities calculated (Table 10). We first sample the mother/proband diplotype combination, and then sample the father's haplotype pairs given the mother/proband haplotype pairs.

[TABLE 9 ABOUT HERE]

[TABLE 10 ABOUT HERE]

3 Results

We validated our case-parent simulation approach using interaction models of sizes one through six (Table 11), for fifteen haplotype blocks containing forty-five SNPs. In each setting, the loci contributing to the genetic risk were in separate haplotype blocks. We chose the haplotype frequencies such that about 5% of the population were carriers of the risk genotype combination. For each setting, we varied the risk among the non-carriers using $\alpha = -5$ (0.7%), $\alpha = -3$ (4.7%), $\alpha = -1$ (27%). We also altered the odds ratios in the risk model (equation 3) using $\beta = 0$ (OR=1), $\beta = 1$ (OR=2.7), $\beta = 2$ (OR=7.4), $\beta = 3$ (OR=20). These extreme values were chosen deliberately, as the objective was to validate the trio simulations. We simulated one hundred data sets with one thousand trios for each α/β combination. It is noteworthy that it is possible to enumerate the complete mating tables, e. g. the trio haplotype pairs and the respective sampling probabilities, only for very limited interaction terms. With this approach, trios under only the first three risk group definitions (Table 11) could be simulated. For the other settings, this approached failed due to excess memory requirements ($>$ 32 GB), and the previously described efficient simulation approach had to be employed.

[TABLE 11 ABOUT HERE]

For each of the simulated data sets, we derived the pseudo-controls as in the genotypic TDT at each of the loci that affected the risk (between one and six loci, see Table 11). Since these loci were chosen in separate blocks, we combined the three pseudo-genotypes in random order at each locus into three pseudo-controls. For all cases and controls we then calculated the Boolean genotype combination that defined risk for each of the cases and pseudo-controls (thus, defining carriers and non-carriers), and used conditional logistic regression with the carrier status as the predictor of interest.

The validation of the trio simulation method was primarily based on the expected values of the parameter estimates derived from the simulated data sets. However, when using conditional logistic regression to compare cases and pseudo-controls, the expected value of the parameter estimates is not the logs odds ratio β, but the log relative risk (Schaid, 1996). In our case, the relative risk is given as

$$RR = \frac{P(D|I_G = 1)}{P(D|I_G = 0)} = \frac{\exp(\alpha+\beta)/(1+\exp(\alpha+\beta))}{\exp(\alpha)/(1+\exp(\alpha))} = \exp(\beta) \times \left(\frac{1+\exp(\alpha+\beta)}{1+\exp(\alpha)}\right)^{-1} \quad (8)$$

and therefore the log relative risk is

$$\log(RR) = \beta - \log\left(\frac{1+\exp(\alpha+\beta)}{1+\exp(\alpha)}\right). \quad (9)$$

The latter term describes the deviation from the logs odds β, and is zero only if β is zero (i. e. risk independent of genotypes), and diminishes as α gets small for $\beta \neq 0$. Notice though that in particular for $\alpha = -1$ in our simulation, the difference between the log relative risk and the log odds ratio can be substantial (Figure 1).

[FIGURE 1 ABOUT HERE]

We also notice that, as expected, the haplotype frequencies in the blocks that carry risk information deviate from the population frequencies (Figure 2) when generating case-parent trios. We observed that the haplotypes that contribute to disease risk were sampled more frequently in our simulations compared to the population at large, at the expense of the other haplotypes in the respective blocks.

[FIGURE 2 ABOUT HERE]

4 Discussion

In this manuscript we presented an extension to the logic regression methodology to detect and assess higher order interactions in trios with affected probands. Trio logic regression accounts for the linkage disequilibrium (LD) structure in the genotype data, and accommodates missing genotypes via haplotype-based imputation. While several approaches to assess SNP-SNP interactions in family data are available (and in particular for nuclear families and affected proband trios), they typically quantify the statistical significance of candidate interactions, and do not allow for the actual interaction search. Trio logic regression is unique in that it employs a conditional logistic regression framework to search and evaluate higher order SNP-SNP interactions, in a non-greedy way. In addition, we also devised an efficient algorithm to simulate case-parent trios where the genetic risk is determined via higher-order epistatic interactions. We validated the algorithm using interactions involving up to 6 SNPs, which, due to computational constraints (and in particular, due to memory constraints), can not be achieved by standard mating table computations.

The efficiency of the simulation approach is in part owed to the fact that only two risk groups are assumed to exist in the population (carriers and non-carriers), which is used for the calculation of the haplotype frequencies in the blocks that carry information about the disease risk. Thus, only one logic tree is permitted in the current trio logic regression. This is a limitation in the explored model space compared to the original logic regression framework introduced by Ruczinski et al. (2003). As described in equation (1) and implemented in the R package `LogicReg`, logic regression in general allows for more than one Boolean term (even though in the vast majority of previously conducted analyses, a single term proved to be sufficient). The single tree assumption in trio logic regression could be relaxed, however, this would make the approach and calculations substantially more complicated. This is in particular the case for the simulation algorithm since many more strata needed to be defined, but also for the actual trio logic regression methodology and software, as for example the criteria to assure convergence had to be augmented (see Appendix). That said, open source software is available to generate a data frame suitable as input for the logic regression R package. If run with a conditional logistic link, more than one tree could be allowed in principle, and might be a future undertaking.

An imperfection is in the haplotype-based imputation employed. Many algorithms and software packages exist to delineate the haplotypes and their frequencies, for both population based and family based designs. However, we expect the haplotype frequencies to differ between cases and controls in disease related blocks, and thus, it might be beneficial to estimate the haplotype frequencies with the algorithm of choice separately for cases and controls. Usually, most SNPs and therefore most LD blocks are not disease related however, and thus, the aforementioned procedure could introduce biases in the analysis simply due to imperfect haplotype frequency estimation in the null SNPs, particularly in settings with small sample sizes. Since it is a priori not known which blocks contain disease related information, and the latter concern seems more severe to us, we do not attempt to estimate haplotype frequencies separately for cases and controls, but only use the ones derived from the founders' genotypes for imputation, to be conservative. It seems reasonable to assume that the loss of information will not be too severe if few data are missing.

We also note that due to recent technological advancements, a major shift from candidate to genome wide association studies (GWAs) has occured. Case-control study designs have arguably been the most popular approach for SNP association studies, and might be even more so in the GWAs settings. Nonetheless, family based GWAs and even GWAs using case-parent designs are carried out (for example, the International Consortium to Identify Genes and Interactions Controlling Oral Clefts, http://www.genevastudy.org/). Obviously, trio logic regression is not suited to investigate interactions between the entire set of SNPs interogated in genome wide association studies. While investigating all 2-way interactions is computationally feasible (e. g. Purcell et al., 2007), searching for higher order interactions is not, due to the vastness of the search space, in particular when probabilistic search algorithms are employed. When permutation tests for model selection are carried out, even data from a custom panel such the Illumina Golden Gate platform (with up to 1536 SNPs) will be computationally challenging. As in other instances when computationally demanding algorithms are to be used, a pre-selection of SNPs might be necessary.

Acknowledgments

Support was provided by NIH grants R01 DK061662 from the National Institute of Diabetes, Digestive and Kidney Diseases, R01 HL090577 from the National Heart, Lung, and Blood Institute, and a CTSA grant to the Johns Hopkins Medical Institutions.

Appendix

The exposures of probands and pseudo-controls are binary, and thus, we can summarize the data by considering, across trios, how many pseudo-controls' exposures equal that of the respective proband, separately for the probands with exposures 0 and 1 (Table 12, upper part).

[TABLE 12 ABOUT HERE]

Considering the respective contributions (Table 12, lower part) to the conditional logistic likelihood in equation (2), we can re-write the log-likelihood as follows:

$$\begin{aligned} & \log(L(\beta)) \\ = \; & a_1 \{\beta - \log(\exp(\beta) + 3)\} + b_1 \{\beta - \log(2\exp(\beta) + 2)\} + c_1 \{\beta - \log(3\exp(\beta) + 1)\} - \\ & a_0 \log(1 + 3\exp(\beta)) - b_0 \log(2 + 2\exp(\beta)) - c_0 \log(3 + \exp(\beta)) + \frac{d_1}{4} + \frac{d_0}{4} \qquad (10) \end{aligned}$$

The first derivative of the log-likelihood yields

$$\begin{aligned}\frac{\partial \log(L(\beta))}{\partial \beta} &= a_1 - (a_1 + c_0) \times \frac{\exp(\beta)}{\exp(\beta) + 3} + b_1 - (b_1 + b_0) \times \frac{2\exp(\beta)}{2\exp(\beta) + 2} + \\ & c_1 - (c_1 + a_0) \times \frac{3\exp(\beta)}{3\exp(\beta) + 1} \qquad (11)\end{aligned}$$

The log-likelihood in equation (11) is monotonically decreasing in β. Further,

$$\lim_{\beta \to -\infty} \frac{\partial \log(L(\beta))}{\partial \beta} = a_1 + b_1 + c_1 \geq 0 \quad \text{and} \quad \lim_{\beta \to +\infty} \frac{\partial \log(L(\beta))}{\partial \beta} = -(a_0 + b_0 + c_0) \leq 0 \quad (12)$$

Since the derivative of the log likelihood function is a continuous function in β, it follows that the likelihood function has a maximum unless $a_1 + b_1 + c_1 = 0$ or $a_0 + b_0 + c_0 = 0$.

References

Andrew AS, Karagas MR, Nelson HH, Guarrera S, Polidoro S, Gamberini S, Sacerdote C, Moore JH, Kelsey KT, Demidenko E, Vineis P, Matullo G (2008) Dna repair polymorphisms modify bladder cancer risk: a multi-factor analytic strategy. Hum Hered 65(2):105–118

Baksh MF, Balding DJ, Vyse TJ, Whittaker JC (2006) A likelihood ratio approach to family-based association studies with covariates. Ann Hum Genet 70(Pt 1):131–139

Baksh MF, Balding DJ, Vyse TJ, Whittaker JC (2007) Family-based association analysis with ordered categorical phenotypes, covariates and interactions. Genet Epidemiol 31(1):1–8

Bhat A, Lucek PR, Ott J (1999) Analysis of complex traits using neural networks. Genet Epidemiol 17 Suppl 1:S503–S507

Breiman L (2001) Statistical modeling: The two cultures. Statistical Science 16(3):199–215

Breslow NE, Day NE, Halvorsen KT, Prentice RL, Sabai C (1978) Estimation of multiple relative risk functions in matched case-control studies. Am J Epidemiol 108(4):299–307

Bureau A, Dupuis J, Falls K, Lunetta KL, Hayward B, Keith TP, Eerdewegh PV (2005) Identifying snps predictive of phenotype using random forests. Genet Epidemiol 28(2):171–182

Chen CH, Chang CJ, Yang WS, Chen CL, Fann CSJ (2003) A genome-wide scan using tree-based association analysis for candidate loci related to fasting plasma glucose levels. BMC Genet 4 Suppl 1:S65

Chen X, Liu CT, Zhang M, Zhang H (2007) A forest-based approach to identifying gene and gene gene interactions. Proc Natl Acad Sci U S A 104(49):19199–19203

Clark TG, Iorio MD, Griffiths RC (2007) Bayesian logistic regression using a perfect phylogeny. Biostatistics 8(1):32–52

Cook NR, Zee RYL, Ridker PM (2004) Tree and spline based association analysis of gene-gene interaction models for ischemic stroke. Stat Med 23(9):1439–1453

Culverhouse R, Hinrichs AL, Jin CH, Suarez BK (2007) Gene x gene and gene x environment interactions for complex disorders. BMC Proc 1 Suppl 1:S72

Culverhouse R, Klein T, Shannon W (2004) Detecting epistatic interactions contributing to quantitative traits. Genet Epidemiol 27(2):141–152

Culverhouse R, Suarez BK, Lin J, Reich T (2002) A perspective on epistasis: limits of models displaying no main effect. Am J Hum Genet 70(2):461–471

Enquobahrie DA, Smith NL, Bis JC, Carty CL, Rice KM, Lumley T, Hindorff LA, Lemaitre RN, Williams MA, Siscovick DS, Heckbert SR, Psaty BM (2008) Cholesterol ester transfer protein, interleukin-8, peroxisome proliferator activator receptor alpha, and toll-like receptor 4 genetic variations and risk of incident nonfatal myocardial infarction and ischemic stroke. Am J Cardiol 101(12):1683–1688

Etzioni R, Falcon S, Gann PH, Kooperberg CL, Penson DF, Stampfer MJ (2004) Prostate-specific antigen and free prostate-specific antigen in the early detection of prostate cancer: do combination tests improve detection? Cancer Epidemiol Biomarkers Prev 13(10):1640–1645

Fallin D, Beaty T, Liang KY, Chen W (2002) Power comparisons for genotypic vs. allelic TDT methods with 2+ alleles. Genet Epidemiol 23(4):458–61; author reply 462–4

Feng Q, Balasubramanian A, Hawes SE, Toure P, Sow PS, Dem A, Dembele B, Critchlow CW, Xi L, Lu H, McIntosh MW, Young AM, Kiviat NB (2005) Detection of hypermethylated genes in women with and without cervical neoplasia. J Natl Cancer Inst 97(4):273–282

Gauderman WJ, Witte JS, Thomas DC (1999) Family-based association studies. J Natl Cancer Inst Monogr (26):31–37

Hahn LW, Ritchie MD, Moore JH (2003) Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. Bioinformatics 19(3):376–382

Harth V, Schafer M, Abel J, Maintz L, Neuhaus T, Besuden M, Primke R, Wilkesmann A, Thier R, Vetter H, Ko YD, Bruning T, Bolt HM, Ickstadt K (2008) Head and neck squamous-cell cancer and its association with polymorphic enzymes of xenobiotic metabolism and repair. J Toxicol Environ Health A 71(13-14):887–897

Heidema AG, Boer JMA, Nagelkerke N, Mariman ECM, van der A DL, Feskens EJM (2006) The challenge for genetic epidemiologists: how to analyze large numbers of snps in relation to complex diseases. BMC Genet 7:23

Huang J, Lin A, Narasimhan B, Quertermous T, Hsiung CA, Ho LT, Grove JS, Olivier M, Ranade K, Risch NJ, Olshen RA (2004) Tree-structured supervised learning and the genetics of hypertension. Proc Natl Acad Sci U S A 101(29):10529–10534

Justenhoven C, Hamann U, Schubert F, Zapatka M, Pierl CB, Rabstein S, Selinski S, Mueller T, Ickstadt K, Gilbert M, Ko YD, Baisch C, Pesch B, Harth V, Bolt HM, Vollmert C, Illig T, Eils

R, Dippon J, Brauch H (2008) Breast cancer: a candidate gene approach across the estrogen metabolic pathway. Breast Cancer Res Treat 108(1):137–149

Keles S, van der Laan MJ, Vulpe C (2004) Regulatory motif finding by logic regression. Bioinformatics 20(16):2799–2811

Kooperberg C, Ruczinski I (2005) Identifying interacting snps using monte carlo logic regression. Genet Epidemiol 28(2):157–170

Kooperberg C, Ruczinski I, LeBlanc ML, Hsu L (2001) Sequence analysis using logic regression. Genet Epidemiol 21 Suppl 1:S626–S631

Kotti S, Bickeboller H, Clerget-Darpoux F (2007a) Strategy for detecting susceptibility genes with weak or no marginal effect. Hum Hered 63(2):85–92

Kotti S, Bourgey M, Clerget-Darpoux F (2007b) Power of the 2-locus tdt for testing the interaction of two susceptibility genes. BMC Proc 1 Suppl 1:S65

Laird NM, Lange C (2006) Family-based designs in the age of large-scale gene-association studies. Nat Rev Genet 7(5):385–394

Laird NM, Lange C (2008) Family-based methods for linkage and association analysis. Adv Genet 60:219–252

Lanktree MB, VanderBeek L, Macciardi FM, Kennedy JL (2004) Pedsplit: pedigree management for stratified analysis. Bioinformatics 20(14):2315–2316

Lucek PR, Ott J (1997) Neural network analysis of complex traits. Genet Epidemiol 14(6):1101–1106

Lunetta KL, Faraone SV, Biederman J, Laird NM (2000) Family-based tests of association and linkage that use unaffected sibs, covariates, and interactions. Am J Hum Genet 66(2):605–614

Lunetta KL, Hayward LB, Segal J, Eerdewegh PV (2004) Screening large-scale association study data: exploiting interactions using random forests. BMC Genet 5(1):32

Martin ER, Bass MP, Gilbert JR, Pericak-Vance MA, Hauser ER (2003) Genotype-based association test for general pedigrees: the genotype-pdt. Genet Epidemiol 25(3):203–213

Martin ER, Ritchie MD, Hahn L, Kang S, Moore JH (2006) A novel method to identify gene-gene effects in nuclear families: the mdr-pdt. Genet Epidemiol 30(2):111–123

McKinney BA, Reif DM, Ritchie MD, Moore JH (2006) Machine learning for detecting gene-gene interactions: a review. Appl Bioinformatics 5(2):77–88

Moore JH (2004) Computational analysis of gene-gene interactions using multifactor dimensionality reduction. Expert Rev Mol Diagn 4(6):795–803

Musani SK, Shriner D, Liu N, Feng R, Coffey CS, Yi N, Tiwari HK, Allison DB (2007) Detection of gene x gene interactions in genome-wide association studies of human population data. Hum Hered 63(2):67–84

North BV, Curtis D, Cassell PG, Hitman GA, Sham PC (2003) Assessing optimal neural network architecture for identifying disease-associated multi-marker genotypes using a permutation test, and application to calpain 10 polymorphisms associated with diabetes. Ann Hum Genet 67(Pt 4):348–356

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007) Plink: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559–575

Ritchie MD (2005) Bioinformatics approaches for detecting gene-gene and gene-environment interactions in studies of human disease. Neurosurg Focus 19(4):E2

Ritchie MD, Hahn LW, Moore JH (2003a) Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. Genet Epidemiol 24(2):150–157

Ritchie MD, Motsinger AA (2005) Multifactor dimensionality reduction for detecting gene-gene and gene-environment interactions in pharmacogenomics studies. Pharmacogenomics 6(8):823–834

Ritchie MD, White BC, Parker JS, Hahn LW, Moore JH (2003b) Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases. BMC Bioinformatics 4:28

Rubin DB (1996) Multiple imputation after 18+ years. Journal of the American Statistical Association 91(434):473–489

Ruczinski I, Kooperberg C, LeBlanc M (2003) Logic regression. Journal of Computational and Graphical Statistics 12(3):475–511

Ruczinski I, Kooperberg C, LeBlanc ML (2004) Exploring interactions in high dimensional genomic data: An overview of logic regression, with applications. Journal of Multivariate Analysis 90:178–95

Schafer JL (1999) Multiple imputation: a primer. Stat Methods Med Res 8(1):3–15

Schaid DJ (1996) General score tests for associations of genetic markers with disease using cases and their parents. Genet Epidemiol 13(5):423–449

Schaid DJ (1999) Likelihoods and tdt for the case-parents design. Genet Epidemiol 16(3):250–260

Schwender H, Ickstadt K (2008) Identification of snp interactions using logic regression. Biostatistics 9(1):187–198

Segal MR, Barbour JD, Grant RM (2004) Relating hiv-1 sequence variation to replication capacity via trees and forests. Stat Appl Genet Mol Biol 3:Article2; discussion article 7, article 9

Self SG, Longton G, Kopecky KJ, Liang KY (1991) On estimating hla/disease association with application to a study of aplastic anemia. Biometrics 47(1):53–61

Spielman RS, Ewens WJ (1996) The tdt and other family-based tests for linkage disequilibrium and association. Am J Hum Genet 59(5):983–989

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (iddm). Am J Hum Genet 52(3):506–516

Suehiro Y, Wong CW, Chirieac LR, Kondo Y, Shen L, Webb CR, Chan YW, Chan ASY, Chan TL, Wu TT, Rashid A, Hamanaka Y, Hinoda Y, Shannon RL, Wang X, Morris J, Issa JPJ, Yuen ST, Leung SY, Hamilton SR (2008) Epigenetic-genetic interactions in the apc/wnt, ras/raf, and p53 pathways in colorectal carcinoma. Clin Cancer Res 14(9):2560–2569

Tomita Y, Tomida S, Hasegawa Y, Suzuki Y, Shirakawa T, Kobayashi T, Honda H (2004) Artificial neural network approach for selection of susceptible single nucleotide polymorphisms and construction of prediction model on childhood allergic asthma. BMC Bioinformatics 5:120

Vaidya VS, Waikar SS, Ferguson MA, Collings FB, Sunderland K, Gioules C, Bradwin G, Matsouaka R, Betensky RA, Curhan GC, Bonventre JV (2008) Urinary biomarkers for sensitive and specific detection of acute kidney injury in humans. Clin Transl Sci 1(3):200–208

Vermeulen SHHM, Heijer MD, Sham P, Knight J (2007) Application of multi-locus analytical methods to identify interacting loci in case-control studies. Ann Hum Genet 71(Pt 5):689–700

Weinberg CR, Wilcox AJ, Lie RT (1998) A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. Am J Hum Genet 62(4):969–978

Yaziji H, Battifora H, Barry TS, Hwang HC, Bacchi CE, McIntosh MW, Kussick SJ, Gown AM (2006) Evaluation of 12 antibodies for distinguishing epithelioid mesothelioma from adenocarcinoma: identification of a three-antibody immunohistochemical panel with maximal sensitivity and specificity. Mod Pathol 19(4):514–523

Zhang Z, Zhang S, Wong MY, Wareham NJ, Sha Q (2008) An ensemble learning approach jointly modeling main and interaction effects in genetic association studies. Genet Epidemiol 32(4):285–300

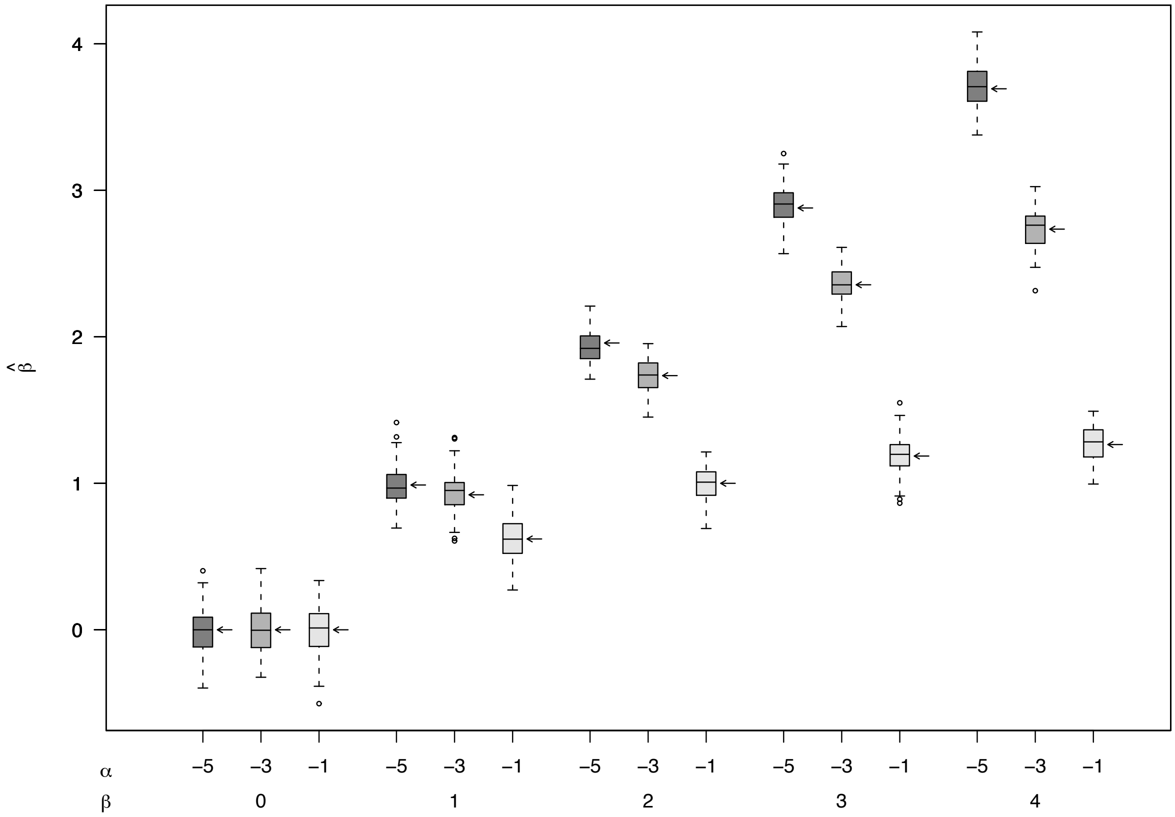

Figure 1: One hundred replicates for 1,000 trios were simulated assuming a risk genotype given by the six-way interaction in Table 11, using various combinations for the parameters α $(-5, -3, -1)$ and β $(0, 1, 2, 3, 4)$. The exact procedure is described in more detail in the text. The boxplots summarize the 100 parameter estimates obtained by using the true risk model as binary predictor in a conditional logistic regression model. The arrows indicate the expected value for the parameters (the log relative risk) as defined in equation (9). The median of each of the parameter estimate sets (shown as a horizontal bar in the center of each box) coincides well with the expected value, thus validating the trio simulation approach. Only the outcome for the six-way interaction is shown. Results and figures for the other six scenarios as indicated in Table 11 were identical. Note that the generation of the mating tables is prohibitive for all but the most simple set-ups (examples 1-3 in Table 11), and a more efficient approach such as the one in the Methods section has to be employed.

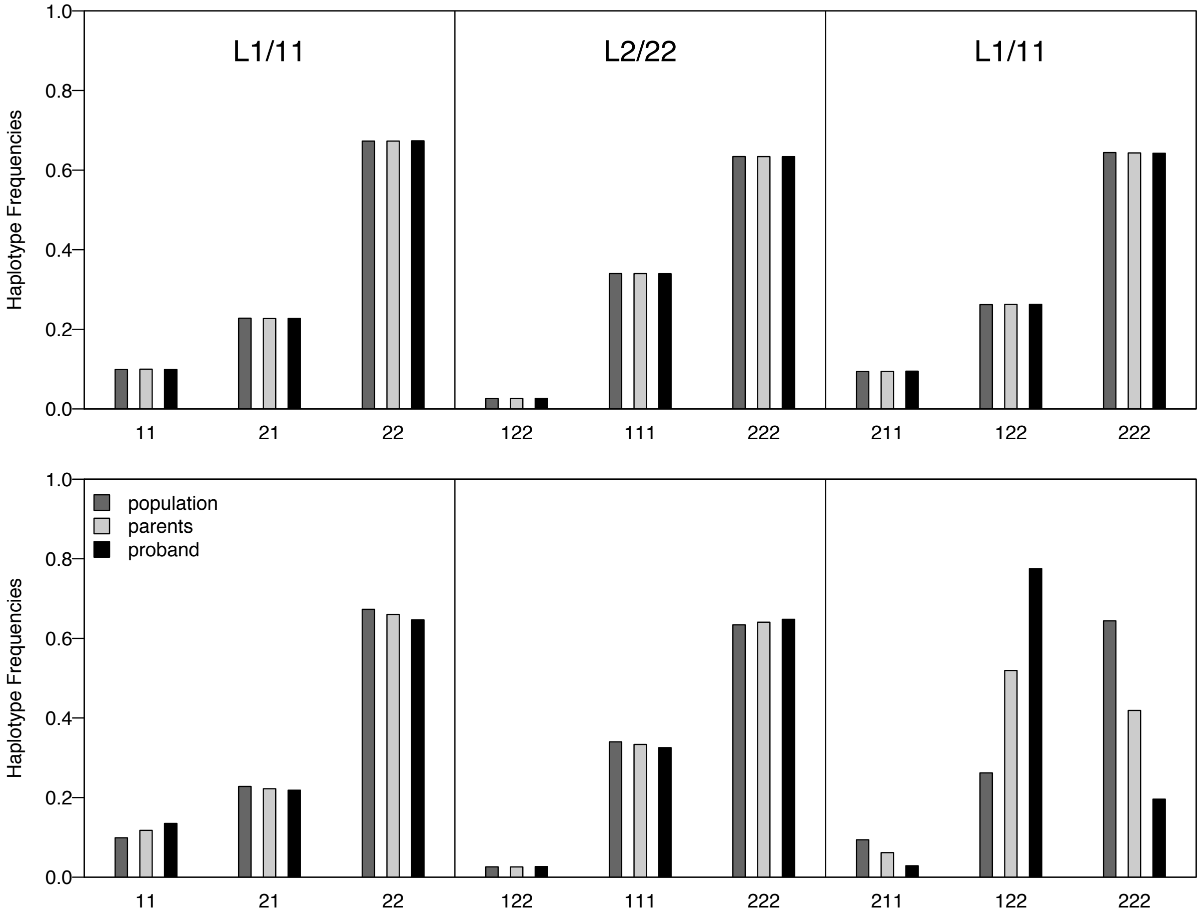

Figure 2: One hundred replicates for 1,000 trios were simulated assuming a risk genotype given by the three-way interaction $(\mathsf{SNP}^R_{1|1} \wedge \overline{\mathsf{SNP}^D_{6|2}}) \vee \mathsf{SNP}^R_{13|1}$ (Table 11, scenario 3), using $\alpha = -5$ and $\beta = 0$ (upper panel) and $\beta = 4$ (lower panel). The above bars show the haplotype frequencies in the three blocks that carry disease risk information (block 1 left, block 6 middle, block 13 right). In each block, we assumed that three different haplotypes exist in the population, as indicated on the horizontal axis. In the above, 1 refers to the minor (variant) allele, and 2 refers to the major allele. Thus, individuals with 2 variant alleles at locus 1 in block 13 (L1/11, right) are at higher risk for disease if $\beta > 0$, as are subjects that have 2 variant alleles at locus 1 in block 1 (L1/11, left) and at the same time no variant alleles at locus 2 in block 6 (L2/22, middle). For each haplotype in each block, we show the population frequency (medium grey), the frequency among the parents (light grey), and the frequency among the probands (dark grey). As expected, if there is no association between genotypes and outcome ($\beta = 0$, upper panel), the haplotype frequencies do not differ among parents, offspring, and the population at large. However, we observe vast differences when such an association exists ($\beta = 4$, lower panel). As expected, the only haplotype that can give rise to genotype 11 at locus 1 in block 13 is 122, and is greatly enriched in parents and particularly in probands (lower right). The same effect, albeit much less pronounced, is still visible for haplotype 11 in block 1 (lower left), and haplotypes 122 and 222 in block 6 (lower middle).

Block	Index	Haplotype	Frequency
1	1	11	$f_{1\|1}$
1	2	21	$f_{2\|1}$
1	3	22	$f_{3\|1}$
2	1	121	$f_{1\|2}$
2	2	122	$f_{2\|2}$
2	3	111	$f_{3\|2}$
2	4	222	$f_{4\|2}$

Table 1: An example of two haplotype blocks having lengths two and three, with three and four possible haplotypes respectively. In the above notation, $f_{i|j}$ refers to the i^{th} haplotype in bock j.

Index	Haplotype index		Haplotype		Frequency
	Block 1	Block 2	Block 1	Block 2	
1	1	1	11	121	$P(H_1) = f_{1\|1} \times f_{1\|2}$
2	2	1	21	121	$P(H_2) = f_{2\|1} \times f_{1\|2}$
3	3	1	22	121	$P(H_3) = f_{3\|1} \times f_{1\|2}$
4	1	2	11	122	$P(H_4) = f_{1\|1} \times f_{2\|2}$
5	2	2	21	122	$P(H_5) = f_{2\|1} \times f_{2\|2}$
6	3	2	22	122	$P(H_6) = f_{3\|1} \times f_{2\|2}$
7	1	3	11	111	$P(H_7) = f_{1\|1} \times f_{3\|2}$
8	2	3	21	111	$P(H_8) = f_{2\|1} \times f_{3\|2}$
9	3	3	22	111	$P(H_9) = f_{3\|1} \times f_{3\|2}$
10	1	4	11	222	$P(H_{10}) = f_{1\|1} \times f_{4\|2}$
11	2	4	21	222	$P(H_{11}) = f_{2\|1} \times f_{4\|2}$
12	3	4	22	222	$P(H_{12}) = f_{3\|1} \times f_{4\|2}$

Table 2: The twelve possible haplotypes across two blocks with three and four haplotypes respectively. The haplotype frequencies are derived from the frequencies in Table 1.

Pair	Haplotype 1	Haplotype 2	genotype
(1, 2)	11121	21121	12-11-11-22-11
(1, 3)	11121	22121	12-12-11-22-11
(1, 7)	11121	11111	11-11-11-12-11
(1, 8)	11121	21111	12-11-11-12-11
(1, 9)	11121	22111	12-12-11-12-11
(2, 3)	21121	22121	22-12-11-22-11
(2, 7)	21121	11111	12-11-11-12-11
(2, 8)	21121	21111	22-11-11-12-11
(2, 9)	21121	22111	22-12-11-12-11
(3, 7)	22121	11111	12-12-11-12-11
(3, 8)	22121	21111	22-12-11-12-11
(3, 9)	22121	22111	22-22-11-12-11
(7, 8)	11111	21111	12-11-11-11-11
(7, 9)	11111	22111	12-12-11-11-11
(8, 9)	21111	22111	22-12-11-11-11
(1, 1)	11121	11121	11-11-11-22-11
(2, 2)	21121	21121	22-11-11-22-11
(3, 3)	22121	22121	22-22-11-22-11
(7, 7)	11111	11111	11-11-11-11-11
(8, 8)	21111	21111	22-11-11-11-11
(9, 9)	22111	22111	22-22-11-11-11
(10, 10)	11222	11222	11-11-22-22-22

Table 3: The diplotypes and genotypes for the risk group, assuming the defining interaction is $(\mathsf{SNP}_1^R \wedge \overline{\mathsf{SNP}_3^D}) \vee \mathsf{SNP}_5^R$. Here, 1 indicates the minor (variant) allele, and 2 indicates the major allele. Therefore, subjects with two variant alleles at locus 5 (genotype 11) are at higher risk, as are subjects with both two variant alleles (genotype 11) at locus 1 and no variant alleles (genotype 22) at locus 3.

Stratum	Risk	Pairs	Index i	Index j	Probability
HR_1	High	Different	1	j $\in \{2,3,7,8,9\}$	$P(H_1)(P(H_2)+P(H_3)+\ldots+P(H_9))$
			2	j $\in \{1,3,7,8,9\}$	$P(H_2)(P(H_1)+P(H_3)+\ldots+P(H_9))$
			3	j $\in \{1,2,7,8,9\}$	$P(H_3)(P(H_1)+P(H_2)+\ldots+P(H_9))$
			7	j $\in \{1,2,3,8,9\}$	$P(H_7)(P(H_1)+P(H_2)+\ldots+P(H_9))$
			8	j $\in \{1,2,3,7,9\}$	$P(H_8)(P(H_1)+P(H_2)+\ldots+P(H_9))$
			9	j $\in \{1,2,3,7,8\}$	$P(H_9)(P(H_1)+P(H_2)+\ldots+P(H_8))$
HR_2	High	Same	1	1	$P(H_1)^2$
			2	2	$P(H_2)^2$
			3	3	$P(H_3)^2$
			7	7	$P(H_7)^2$
			8	8	$P(H_8)^2$
			9	9	$P(H_9)^2$
			10	10	$P(H_{10})^2$
LR_1	Low	Same	4	4	$P(H_4)^2$
			...	...	...
			12	12	$P(H_{12})^2$
LR_2	Low	Different	1	j $\notin \{1,2,3,7,8,9\}$	$P(H_1)(1-P(H_1)-P(H_2)-\ldots-P(H_9))$
			2	j $\notin \{1,2,3,7,8,9\}$	$P(H_2)(1-P(H_1)-P(H_2)-\ldots-P(H_9))$
			...	...	...
			11	$\neq 11$	$P(H_{11})(1-P(H_{11}))$
			12	$\neq 12$	$P(H_{12})(1-P(H_{12}))$

Table 4: The haplotype pairs in a population are tabulated according to high and low-risk diplotypes, as given in Table 3. For computational efficiency and convenience, we also differentiate between identical and non-identical haplotypes in a diplotype, denoted as strata HR_1, HR_2, LR_1, and LR_2, and further stratify by the possible indices of the first haplotype we will sample in a pair (Index i). The strata probabilities are derived from the frequencies in Table 2. Note that this somewhat time-consuming tabulation step is carried out when a simulation is initialized, and thus, has to be invoked only once.

Strata	Index 1	Index 2	Probability
Child with different haplotypes (i, j)			
Parent 1			
A_1	i	$\neq i$, $\neq j$	$2P(H_i)(1 - P(H_i) - P(H_j))$
A_2	i	j	$2P(H_i)P(H_j)$
A_3	i	i	$P(H_i)^2$
Total →			$2P(H_i) - P(H_i)^2$
Parent 2			
B_1	j	$\neq i$, $\neq j$	$2P(H_j)(1 - P(H_i) - P(H_j))$
B_2	j	i	$2P(H_i)P(H_j)$
B_3	j	j	$P(H_j)^2$
Total →			$2P(H_j) - P(H_j)^2$
Child with identical haplotypes (i, i)			
Any Parent			
C_1	i	$\neq i$	$2P(H_i)(1 - P(H_i))$
C_2	i	i	$P(H_i)^2$
Total →			$2P(H_i) - P(H_i)^2$

Table 5: The distributions for the parental haplotype pairs assuming the child has non-identical (i, j) or identical (i, i) haplotypes. The strata probabilities are derived from the frequencies in Table 2.

Index	Parent 1	Parent 2	Factor P	Factor C	Sampling probability
Child with different haplotypes (i, j)					
1	A_1: (i, k)	B_1: (j, k)	2	$\frac{1}{4}$	$\frac{1}{2}P(A_1)P(B_1)$
2	A_1: (i, k)	B_2: (i, j)	2	$\frac{1}{4}$	$\frac{1}{2}P(A_1)P(B_2)$
3	A_1: (i, k)	B_3: (j, j)	2	$\frac{1}{2}$	$P(A_1)P(B_3)$
4	A_2: (i, j)	B_1: (j, k)	2	$\frac{1}{4}$	$\frac{1}{2}P(A_2)P(B_1)$
5	A_2: (i, j)	B_2: (i, j)	1	$\frac{1}{2}$	$\frac{1}{2}P(A_2)P(B_2)$
6	A_2: (i, j)	B_3: (j, j)	2	$\frac{1}{2}$	$P(A_2)P(B_3)$
7	A_3: (i, i)	B_1: (j, k)	2	$\frac{1}{2}$	$P(A_3)P(B_1)$
8	A_3: (i, i)	B_2: (i, j)	2	$\frac{1}{2}$	$P(A_3)P(B_2)$
9	A_3: (i, i)	B_3: (j, j)	2	1	$2P(A_3)P(B_3)$
Total →					$2P(H_i)P(H_j)$
Child with identical haplotypes (i, i)					
1	C_1: (i, j)	C_1: (i, k)	1	$\frac{1}{4}$	$\frac{1}{4}P(C_1)^2$
2	C_2: (i, i)	C_1: (i, k)	2	$\frac{1}{2}$	$P(C_1)P(C_2)$
3	C_2: (i, i)	C_2: (i, i)	1	1	$P(C_2)^2$
Total →					$P(H_i)^2$

Table 6: The parental diplotype pair distribution assuming the child has non-identical (i, j) or identical (i, i) haplotypes. The strata probabilities are based on the haplotype pair frequencies in Table 5. Two further considerations are important in the derivation of the sampling probabilities. The joint probability for a non-identical pair of diplotypes in the parents has to be multiplied by the factor 2, since there is no ordering in the parental diplotype pairs (denoted above as *Factor P*). Further, the number of distinct diplotypes in the four "Mendelian" children can be one, two, or four, depending on the parents' haplotype combinations. This is taking into consideration by multiplying the parents' diplotype probabilities with the respective factor (denoted above as *Factor C*).

	Father	Mother	Proband	Sampling probability	Implementation
1	MC	MC	MC	$P(C\|F,M)P(F)P(M)$	Randomly draw the parental diplotypes, then derive the proband's diplotype.
2	CP	CP	MC	$P(C\|F,M)P(F\|F_G)P(M\|M_G)$	Draw parental diplotypes conditioning on the observed genotypes, then derive the proband's diplotype.
3	MC	CP	MC	$P(C\|F,M)P(F)P(M\|M_G)$	Randomly draw the father's diplotype, draw the mother's conditioning on the observed genotypes, then derive the proband's.
	CP	MC	MC	$P(C\|F,M)P(F\|F_G)P(M)$	Randomly draw the mother's diplotype, draw father's conditioning on the observed genotypes, then derive the proband's.
4	MC	MC	CP	$P(F,M\|C)P(C\|C_G)$	Draw the proband's diplotype conditioning on the observed genotypes, then sample the parental diplotypes.
5	MC	CP	CP	$P(F,M,C\|M_G,C_G)$	Draw the diplotype pair for the mother and child, then sample the paternal haplotype pair.
	CP	MC	CP	$P(F,M,C\|F_G,C_G)$	Draw the diplotype pair for the father and child, then sample the maternal haplotype pair.
6	CP	CP	CP	$P(C\|F,M,C_G)P(F,M\|F_G,M_G,C_G)$	Enumerate the possible parental diplotypes, then derive the proband's diplotypes. Eliminate the impossible combinations after comparing the diplotypes with the observed genotypes in the proband.

Table 7: The different strategies to derive the haplotypes and impute the missing data. We distinguish between a subject's genotype being *Missing Completely (MC)* and being *Complete or Partially complete (CP)*. Further, F, M and C abbreviate the father's, mother's and child's haplotype pair. We use the subscript G to distinguish the genotypes from the haplotypes. When the genotypes of only one parent are completely missing and the child's genotypes are partially or completely observed (strata 5), all haplotype pairs have to be considered jointly. However, this can be done in an efficient manner (see the text, and Table 8).

Category	Father		Mother	Proband	Factor	Sampling probabilities
A	i	i	(i, i)	(i, i)	1	$P(H_i)^2 P(H_{Mi})^2$
	i	≠i	(i, i)	(i, i)	$\frac{1}{2}$	$P(H_i)(1 - P(H_i))P(H_{Mi})^2$
B	i	i	(i, k)	(i, k)	$\frac{1}{2}$	$P(H_i)^2 P(H_{Mi})P(H_{Mk})$
	i	≠i, ≠k	(i, k)	(i, k)	$\frac{1}{4}$	$P(H_i)(1 - P(H_i) - P(H_k))P(H_{Mi})P(H_{Mk})$
	i	k	(i, k)	(i, k)	$\frac{1}{2}$	$2P(H_i)P(H_k)P(H_{Mi})P(H_{Mk})$
	k	≠i, ≠k	(i, k)	(i, k)	$\frac{1}{4}$	$P(H_k)(1 - P(H_i) - P(H_k))P(H_{Mi})P(H_{Mk})$
	k	k	(i, k)	(i, k)	$\frac{1}{2}$	$P(H_k)^2 P(H_{Mi})P(H_{Mk})$
C	i	i	(i, j)	(i, i)	$\frac{1}{2}$	$P(H_i)^2 P(H_{Mi})P(H_{Mj})$
	i	≠i, ≠j	(i, j)	(i, i)	$\frac{1}{4}$	$P(H_i)(1 - P(H_i) - P(H_j))P(H_{Mi})P(H_{Mj})$
	i	j	(i, j)	(i, i)	$\frac{1}{4}$	$P(H_i)P(H_j)P(H_{Mi})P(H_{Mj})$
D	k	k	(i, i)	(i, k)	1	$P(H_k)^2 P(H_{Mi})^2$
	k	≠k	(i, i)	(i, k)	$\frac{1}{2}$	$P(H_k)(1 - P(H_k))P(H_{Mi})^2$
E	k	k	(i, j)	(i, k)	$\frac{1}{2}$	$P(H_k)^2 P(H_{Mi})P(H_{Mj})$
	k	≠k	(i, j)	(i, k)	$\frac{1}{4}$	$P(H_k)(1 - P(H_k))P(H_{Mi})P(H_{Mj})$

Table 8: Sampling probabilities for trio haplotype pairs, stratified by mother/child diplotypes, and distinguishing the different instances for the father's transmitted allele. The sampling probabilities are based on the father's haplotypes, the mother's haplotypes subject to the genotype constraints (denoted as standardized haplotype frequencies, $P(H_M)$), and the number of distinct "Mendelian" children, given the parents' diplotypes. This number of distinct offspring diplotypes can be one, two, or four, and is taking into consideration by multiplying the parental probabilities with the respective factor (denoted above as *Factor*).

index	haplotype	frequency
1	111	$P(H_1)$
2	211	$P(H_2)$
3	121	$P(H_3)$
4	221	$P(H_4)$
5	112	$P(H_5)$
6	212	$P(H_6)$
7	122	$P(H_7)$
8	222	$P(H_8)$

Table 9: Index and notation for the haplotypes in a three locus block and the respective haplotype frequencies, used as an example in the text.

	Father		Mother	Proband	Factor	Sampling probabilities
A	8	8	(8, 8)	(8, 8)	1	$P(H_8)^2 P(H_{M8})^2$
	8	$\neq$ 8	(8, 8)	(8, 8)	$\frac{1}{2}$	$P(H_8)(1 - P(H_8))P(H_{M8})^2$
C	8	8	(3, 8)	(8, 8)	$\frac{1}{2}$	$P(H_8)^2 P(H_{M3})P(H_{M8})$
	8	$\neq$3, 8	(3, 8)	(8, 8)	$\frac{1}{4}$	$P(H_8)(1 - P(H_3) - P(H_8))P(H_{M3})P(H_{M8})$
	8	3	(3, 8)	(8, 8)	$\frac{1}{4}$	$P(H_8)P(H_3)P(H_{M3})P(H_{M8})$
	8	8	(4, 8)	(8, 8)	$\frac{1}{2}$	$P(H_8)^2 P(H_{M4})P(H_{M8})$
	8	$\neq$4, 8	(4, 8)	(8, 8)	$\frac{1}{4}$	$P(H_8)(1 - P(H_4) - P(H_8))P(H_{M4})P(H_{M8})$
	8	4	(4, 8)	(8, 8)	$\frac{1}{4}$	$P(H_8)P(H_4)P(H_{M4})P(H_{M8})$
	8	8	(7, 8)	(8, 8)	$\frac{1}{2}$	$P(H_8)^2 P(H_{M7})P(H_{M8})$
	8	$\neq$7, 8	(7, 8)	(8, 8)	$\frac{1}{4}$	$P(H_8)(1 - P(H_7) - P(H_8))P(H_{M7})P(H_{M8})$
	8	7	(7, 8)	(8, 8)	$\frac{1}{4}$	$P(H_8)P(H_7)P(H_{M7})P(H_{M8})$
D	6	6	(8, 8)	(6, 8)	1	$P(H_6)^2 P(H_{M8})^2$
	6	$\neq$6	(8, 8)	(6, 8)	$\frac{1}{2}$	$P(H_6)(1 - P(H_6))P(H_{M8})^2$
E	6	6	(3, 8)	(6, 8)	$\frac{1}{2}$	$P(H_6)^2 P(H_{M3})P(H_{M8})$
	6	$\neq$6	(3, 8)	(6, 8)	$\frac{1}{4}$	$P(H_6)(1 - P(H_6))P(H_{M3})P(H_{M8})$
	6	6	(4, 8)	(6, 8)	$\frac{1}{2}$	$P(H_6)^2 P(H_{M4})P(H_{M8})$
	6	$\neq$6	(4, 8)	(6, 8)	$\frac{1}{4}$	$P(H_6)(1 - P(H_6))P(H_{M4})P(H_{M8})$
	6	6	(7, 8)	(6, 8)	$\frac{1}{2}$	$P(H_6)^2 P(H_{M7})P(H_{M8})$
	6	$\neq$6	(7, 8)	(6, 8)	$\frac{1}{4}$	$P(H_6)(1 - P(H_6))P(H_{M7})P(H_{M8})$

Table 10: The sampling probabilities for the trio haplotype pairs as in Table 8, assuming a three-locus haplotype black and using the notation from Table 9.

	G	$P(G)$	Haplotypes / block	n_1	n_2
1	$\text{SNP}^{R}_{13\mid 1}$	0.069	3	3	21
2	$\text{SNP}^{R}_{4\mid 2} \wedge \text{SNP}^{R}_{5\mid 3}$	0.053	8, 5	40	$\approx 10^{5}$
3	$(\text{SNP}^{R}_{1\mid 1} \wedge \overline{\text{SNP}^{D}_{6\mid 2}}) \vee \text{SNP}^{R}_{13\mid 1}$	0.073	3, 3, 3	27	$\approx 10^{4}$
4	$(\text{SNP}^{R}_{2\mid 3} \wedge \overline{\text{SNP}^{D}_{5\mid 2}}) \vee \text{SNP}^{R}_{7\mid 3}$	0.060	4, 5, 8	160	$\approx 10^{7}$
5	$(\overline{\text{SNP}^{D}_{4\mid 1}} \wedge \text{SNP}^{R}_{8\mid 2}) \vee (\text{SNP}^{R}_{5\mid 1} \wedge \overline{\text{SNP}^{D}_{6\mid 1}})$	0.057	8, 5, 5, 3	600	$\approx 10^{10}$
6	$((\text{SNP}^{R}_{4\mid 2} \vee \text{SNP}^{R}_{7\mid 2}) \wedge \text{SNP}^{R}_{8\mid 3}) \vee (\text{SNP}^{R}_{9\mid 2} \wedge \text{SNP}^{R}_{6\mid 1})$	0.063	8, 8, 5, 3, 3	2,880	$\approx 10^{12}$
7	$(\text{SNP}^{R}_{3\mid 11} \wedge \text{SNP}^{R}_{12\mid 2}) \vee (\text{SNP}^{R}_{5\mid 4} \wedge \text{SNP}^{R}_{15\mid 1}) \vee (\text{SNP}^{R}_{9\mid 2} \wedge \text{SNP}^{R}_{8\mid 3})$	0.066	3, 4, 5, 5, 3, 5	4,500	$\approx 10^{13}$

Table 11: The interactions in the genetic models used to validate the method and algorithm for the case-parent trio simulation. We simulated fifteen haplotype blocks containing forty-five SNPs based on the above interactions, with various parameters for the disease risk model (see text for details). In the above notation, $\text{SNP}^{R}_{13|1}$ is equal to one if the first locus in block 13 has two variant alleles. The superscript D denotes an assumed dominant effect (one or two variant alleles), and the horizontal bar denotes the Boolean complement. The symbols $\vee$ and $\wedge$ stand for the Boolean operators *or* and *and*, respectively. Further, $P(G)$ denotes the proportion of risk carriers in the population. The number of haplotypes in the blocks that contain a locus with disease risk information was chosen between 3 and 8. The total number of possible haplotypes is recorded (n_1), as is the number of rows in the respective mating tables (n_2). The efficient simulation method introduced was crucial for the interactions 4–8 as the memory requirements were prohibitive when attempting to calculate the mating table.

	Number of trios where k pseudo-controls match the proband			
$k \rightarrow$	0	1	2	3
$X_{i0} = 0$	a_0	b_0	c_0	d_0
$X_{i0} = 1$	a_1	b_1	c_1	d_1

	Likelihood contributions for $X_{i0} = 0$			
	a_0	b_0	c_0	d_0
X_{i0}	0	0	0	0
X_{i1}	1	0	0	0
X_{i2}	1	1	0	0
X_{i3}	1	1	1	0
$L_i(\beta)$	$\frac{1}{1+3\exp(\beta)}$	$\frac{1}{2+2\exp(\beta)}$	$\frac{1}{3+\exp(\beta)}$	$\frac{1}{4}$

	Likelihood contributions for $X_{i0} = 1$			
	a_1	b_1	c_1	d_1
X_{i0}	1	1	1	1
X_{i1}	0	1	1	1
X_{i2}	0	0	1	1
X_{i3}	0	0	0	1
$L_i(\beta)$	$\frac{\exp(\beta)}{\exp(\beta)+3}$	$\frac{\exp(\beta)}{2\exp(\beta)+2}$	$\frac{\exp(\beta)}{3\exp(\beta)+1}$	$\frac{1}{4}$

Table 12: The number of exposures of pseudo-controls that equal the exposure of the respective probands, across all trios, separate for the two possibilities for the probands' exposures. For example, a_0 is the number of trios for which the probands exposure is 0 and all pseudo-controls' exposure is equal to one. Each of these trios then contributes $1/(1 + 3\exp(\beta))$ to the likelihood. Note that trios for which the exposures of all pseudo-controls are equal to the exposure of the proband (d_0 and d_1) do not contribute information about β.