Published in final edited form as:

Genet Epidemiol. 2009 January ; 33(1): 54–62. doi:10.1002/gepi.20356.

COMBINED HAPLOTYPE RELATIVE RISK (CHRR):

A GENERAL AND SIMPLE GENETIC ASSOCIATION TEST THAT COMBINES TRIOS AND UNRELATED CASE-CONTROLS

Chao-Yu Guo^{1,2,3}, Kathryn L. Lunetta⁴, Anita L. DeStefano⁴, and L. Adrienne Cupples⁴

¹Clinical Research Program, Children's Hospital Boston

²Program in Genomics, Children's Hospital Boston

³Department of Pediatrics, Harvard Medical School

⁴Department of Biostatistics, Boston University School of Public Health

Abstract

In some genetic association studies, samples contain both parental and unrelated controls. Under such scenarios, instead of analyzing only trios using family-based association tests or only unrelated subjects using a case-control study design, [Nagelkerke et al., 2004] and [Epstein et al., 2005] proposed methods that implemented a likelihood ratio test to combine the two different types of data. In this article, we put forward a more powerful and simplified strategy to combine trios with unrelated subjects based on the Haplotype Relative Risk (HRR) [Falk et al., 1987]. The HRR compares parental marker alleles transmitted to an affected offspring to those not transmitted as a test for association, a strategy that is similar to a case-control study that compares allele frequencies in diseased cases to those of unrelated controls. We prove that affected offspring can be pooled with diseased cases and that parental controls can be treated as unrelated controls when the trios and unrelated subjects are randomly sampled from the same population. Therefore, unrelated subjects can be incorporated into the HRR intuitively and effortlessly. For trios without complete parental genotypes, we adopted the strategy proposed by [Guo et al., 2005], which is more feasible than the one proposed by [Weinberg, 1999]. In addition, simulation results suggest that the CHRR is more powerful than Epstein et al.'s method regardless of the disease prevalence in a homogeneous population.

Keywords

TDT; HRR; case-control association study; trios; missing data

INTRODUCTION

In genetic association studies, there are two popular designs to detect association between genetic markers and phenotypes. The first is the case-control study using unrelated individuals while the second is the family-based transmission disequilibrium design. These two designs require different strategies for collection of subjects to be included in analyses. However, many studies include both biologically related and unrelated individuals. If case-

Corresponding author: Chao-Yu Guo, Ph.D Assistant Professor Clinical Research Program and Program in Genomics Children's Hospital Boston and Harvard Medical School Address: 300 Longwood Avenue, Boston, MA 02115 Phone: 857-218-4718 Fax: 617-730-0828 E-mail: chao-yu.guo @childrens.harvard.edu.

SOFTWARE

The SAS macro for CHRR is available. Please contact Dr. Chao-Yu Guo for the macro at chao-yu.guo@childrens.harvard.edu

control analyses are used, biologically related subjects are excluded and similarly, if family-based TDT analyses are used, unrelated subjects are excluded. This paper proposes a method that combines the information from both types of subjects, when both are available.

The first design we consider is case-control studies. While many researchers prefer the relative risk (RR) as an estimate of increased risk, it requires prospective information. Thus, in case-control studies, the RR is frequently approximated by the odd ratio (OR) given by

the cross product $\frac{\Pr(G|\text{Affected Cases}) \times \Pr(g|\text{Unaffected Controls})}{\Pr(g|\text{Affected Cases}) \times \Pr(G|\text{Unaffected Controls})}$, where G stands for the presence of the reference genotype and g for the lack of the reference genotype. When the overall frequency of the disease in a population is low, the OR closely approximates the relative risk.

Several assumptions are generally made about the underlying population from which both the diseased and control samples are obtained, most importantly that both samples are drawn from the same genetically homogeneous population in an unbiased manner. Therefore, when population admixture exists, spurious association may occur. Several methods have been proposed [Devlin and Roeder 1999; Epstein, et al. 2007; Price, et al. 2006; Pritchard, et al. 2000; Pritchard and Rosenberg 1999; Purcell, et al. 2007] to correct for the bias in statistical tests when admixture exists using genotype information from unlinked markers.

The second design is the family-based study design proposed by [Rubinstein, et al. 1981] and [Falk and Rubinstein 1987] which can avoid spurious association due to population stratification in a traditional case-control study, in which families are ascertained through a single affected offspring. They suggested the use of parental haplotypes present in the affected offspring as cases and those not present in the affected offspring as controls; hence the method was named the haplotype relative risk (HRR). The HRR tests for linkage disequilibrium (LD) between a marker and a putative disease locus by comparing parental marker alleles transmitted to an affected offspring to those not transmitted.

Unlike the HRR method, [Spielman, et al. 1993] viewed the trios in the family-based design as a matched case control study, where the transmitted allele is matched to the non-transmitted for each parent. Hence, each trio generates two observations instead of the four used in the HRR statistic. For this case, the McNemar Test [Sokal and Rohlf 1969] is used to analyze such matched data and the Transmission/Disequilibrium Test (TDT) compares the number of heterozygous parents who transmitted one allele to the number of heterozygous parents who transmitted the other allele to affected offspring. Although the HRR is free from spurious association, it is more conservative and its power decreases more rapidly than that of the TDT with respect to the degree of population admixture [Ewens and Spielman 1995; Guo, et al. 2005a; Guo, et al. 2005b] Therefore, the TDT is preferred in the family-based genetic studies that test for linkage and/or association.

Although family-based trios are robust to population admixture, parental genotypes are often more expensive and difficult to collect than unrelated controls. In addition, the HRR and TDT are not generally used when parental genotypes are incomplete because of death or refusal to participate. Under this scenario, several strategies have been proposed [Allen, et al. 2003; Chen 2004; Clayton 1999; Guo, et al. 2005b; Sebastiani, et al. 2004; Weinberg 1999] to accommodate incomplete trios. Even if parental controls are properly handled, the difficulty of collecting them may result in insufficient sample size to detect the LD between a putative disease gene and marker.

The determination of whether a family-based or an unrelated case-control study should be used requires deliberate consideration, since both study designs provide advantages

compared to the other. The main advantage of the case-control design is that unrelated samples are easier ascertained than related family members. Assuming no admixture, the case-control study provides more power per subject genotyped compare to the family-based design. [Veal, et al. 2002] used case-parent trios for the family-based association analysis and highlighted two SNPs associated with psoriasis. They additionally used a separate set of unrelated controls in order to tease out whether there were two distinct genes contributing to risk or if LD in this region may explain the association to the two genes. [Martin and Kaplan 2000] collected trios to confirm previous association results that were found using a sample of unrelated cases and controls, because the use of trios ensures significant results are due to association and not population stratification. Instead of analyzing either trios or unrelated subjects alone, [Nagelkerke, et al. 2004] discussed the possibility of pooling family trios and unrelated cases and controls using a likelihood-based approach and revealed increased power compared to analyzing trios or unrelated subjects alone. However, their strategy, although easily conducted, used an approximate analysis that requires strong assumptions of Hardy-Weinberg equilibrium (HWE), random mating, and a multiplicative model of allele effect on disease.

Later, [Epstein, et al. 2005] advocated a likelihood-based approach that modifies the approach of [Nagelkerke, et al. 2004] to allow for more flexible modeling of allelic effects and less restrictive assumptions about the distribution of parental mating-types and genotypes. They found that the power gains achieved by [Nagelkerke, et al. 2004] under the assumption of HWE and random mating were preserved when a more general parental mating-type distribution is used, whereas violation of these assumptions in the approximate procedure of [Nagelkerke, et al. 2004] can lead to biased inference. In addition, they provided formal tests to determine whether trios and unrelated cases and controls can be combined, under which case genotyping at additional loci was not required. They also considered trios without complete parental genotypes. However, Epstein et al.'s likelihood approach requires the assumption that the disease is rare.

The motivation of this research is that [Ott 1989] and [Knapp, et al. 1993] discussed that the HRR and family-based association tests can detect marker/disease associations only if the marker and disease genes are in linkage disequilibrium and are also linked, as opposed to case-controls studies which can detect association in the absence of linkage. In addition, [Terwilliger and Ott 1992] showed that when random-mating assumptions hold, a contingency table approach is more powerful than the HRR or TDT. [Thomson 1995] later derived a theoretical derivation of the parental controls frequencies, for bi-allelic and multi-allelic loci and not limited to trios but applicable to various multiplex family structures. Thomson showed that family based data will detect association only if the recombination fraction is less than 0.5, and that even if the recombination fraction is greater than 0, association can still be detected, but the estimates of control frequencies derived from families in this instance will not be an unbiased estimator of population frequencies.

Therefore, previous research has demonstrated the strength and similarity between the HRR and case-control studies. Instead of calculating complex likelihood functions parental mating-type distributions that assume a rare disease condition, we present a general, simple and powerful method called the combined haplotype relative risk (CHRR), which jointly analyzes case-parent trios and unrelated cases and controls based on the HRR. We also adopt the Expectation-Maximization algorithm based haplotype relative risk (EM-HRR) approach proposed by [Guo, et al. 2005b], which is more feasible than the one proposed by [Weinberg, 1999], to accommodate both dyads (affected child with one parental genotype) and monads (affected child without parental genotype). In addition, a logistic regression framework is further proposed to allow the CHRR to adjust for covariates and examine gene-environment interaction.

STATISTICAL METHOD

Following the assumptions in [Epstein, et al. 2005; Nagelkerke, et al. 2004], we consider the underlying population to be homogeneous. In this circumstance, trios and unrelated cases and controls can be combined, but the rare disease assumption is absent in our approach. In the following, we show via theoretical proof that the probability density functions (PDFs) of marker alleles for the affected offspring in the ascertained trios are identical to that of the unrelated affected cases when testing for association in the presence of linkage, if subjects were randomly selected from the same population. In addition, we prove that the PDFs of marker alleles for parental controls and unrelated controls are also identical under the null hypothesis of no association in the presence of linkage. Thus, affected offspring can be pooled with unrelated diseased cases and parental controls can be treated as the unrelated controls under the null hypothesis. Hence, the unrelated subjects are incorporated into the HRR intuitively and effortlessly.

For illustration, we present derivations of the PDFs of marker alleles under a recessive disease model. The PDFs of marker alleles for other disease models can be derived in a similar manner. Let B_1 denote a particular allele at a marker locus with population frequency “ r ”, and B_2 denote any other allele (absence of B_1) with a population frequency “ s ” ($r + s = 1$). At the disease locus, let “ a ” denote the disease-predisposing allele with a population frequency “ q ”, while “ A ” is the normal allele with a population frequency “ p ” ($p + q = 1$) and “ A ” is dominant over “ a ” (recessive disease). Let “ δ ” = $p(aB_1) - p(a)p(B_1)$ denote the disequilibrium coefficient, where $\max\{-(1-q)r, -q(1-r)\} \leq \delta \leq \min\{qr, (1-q)(1-r)\}$. “ $\delta = 0$ ” means no association between the disease and marker alleles. The four haplotype probabilities X_1, X_2, X_3 and X_4 are displayed in Table 1.

Based on the four haplotype probabilities, [Ott 1989] derived the joint distribution functions of transmitted and non-transmitted marker alleles for one parent, as displayed in Table 2, using the previously defined symbols. Since the HRR compares parental marker alleles transmitted to an affected offspring to those not transmitted, the marginal probabilities in Table 2 are used in the HRR and are displayed in the first (affected offspring) and third (parental controls) columns of Table 3.

[Ott 1989] commented that the HRR does not provide a valid χ^2 test for testing linkage and association, because this statistic assumes independence of allelic types of transmitted or non-transmitted alleles for all parents. Such independence will hold if and only if each probability in Table 2 is the product of the corresponding marginal probabilities, and algebraic manipulation shows that this requires $\theta \times \delta = 0$. Without violating this statistical property, one can claim that the HRR is a valid test for the null hypothesis of no association ($\delta = 0$) in the presence of linkage ($\theta < 0.5$).

In the appendix, we calculated the probability density functions for unrelated subjects, which are displayed in the second (unrelated cases) and last (unrelated controls) columns of Table 3. One can see that under the null hypothesis of no association ($\delta = 0$) in the presence of linkage ($\theta < 0.5$):

$$P(B_1 \mid \text{unrelated controls}) = P(B_1 \mid \text{parental controls}) = r$$

$$P(B_2 \mid \text{unrelated controls}) = P(B_2 \mid \text{parental controls}) = 1 - r$$

Therefore, parental and unrelated controls have identical distributions and can be pooled in the analysis. Similarly, affected offspring and unrelated cases can be analyzed jointly, because under the null hypothesis:

$$P(B_1|\text{unrelated cases}) = P(B_1|\text{affected offspring}) = r$$

$$P(B_2|\text{unrelated cases}) = P(B_2|\text{affected offspring}) = 1 - r.$$

Note that the above derivations do not require the assumption that the disease is rare as in [Epstein, et al. 2005]. Distributions under the dominant disease model can be derived in the same manner and are displayed in the appendix (Table A.1).

Based on Table 3, denote the cell frequency in row i ($i = 1, 2$ if the allele is B_1, B_2) and column j by N_{ij} ($j = 1, 3$ if the allele is from affected trios; $j = 2, 4$ if the allele is from case-controls). By collapsing trios and case-controls, a 2×2 table is formed with four cells $A = N_{11} + N_{12}$, $B = N_{13} + N_{14}$, $C = N_{21} + N_{22}$, and $D = N_{23} + N_{24}$.

Therefore, A, B, C , and D represent the total number of the B_1 alleles in affected offspring and unrelated cases, the total number of the B_1 alleles in parental and unrelated controls, the total number of the B_2 alleles in affected offspring and unrelated cases, and the total number of the B_2 alleles in parental and unrelated controls, respectively. The CHRR is defined as

$\frac{A \times D}{B \times C}$. Under the null hypothesis of no association in the presence of linkage, the CHRR is

expected to be 1 and the test statistic $\chi^2_{CHRR} = \frac{[\log(CHRR)]^2}{\text{Var}(\log(CHRR))}$ follows a central chi-square distribution with 1 degree of freedom, where $\text{Var}(\log(CHRR))$ is approximated by

$$\left(\frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D} \right).$$

INCOMPLETE TRIOS

In this section, we adopt the EM-HRR approach proposed by [Guo, et al. 2005b] to accommodate dyads and monads. The HRR compares parental marker alleles transmitted to an affected offspring to those not transmitted. One feature of the HRR is that the affected offspring's genotypes are always known (assuming no genotyping failure) due to the ascertainment criteria, which collects an affected individual first and then seeks his/her parents. Thus, in the case group, the two transmitted alleles of all affected offspring are known and can be used in the analysis, even when both parents' genotypes are not available.

Let $M_k^{i,j}$ represent the observed count for each type of trio data, where $k = "0", "1"$ or $"2"$ represents the total number of B_1 alleles transmitted to the offspring, and $i, j = "0", "1"$ or $"2"$ represents the total number of B_1 alleles for fathers and mothers, respectively. Note that the superscript $"*"$ is used when the parental genotype is missing. Therefore, dyads are denoted by $M_k^{i,*}$ whenever the father or mother is missing, assuming no difference according to the sex of the parent. [Curtis and Sham 1995] showed that bias in estimating the probability of transmission of certain alleles is introduced if incomplete trios with one heterozygous offspring and one heterozygous parent are excluded (exclusion of $M_1^{1,*}$) from the analysis. Later, the use of the EM-algorithm proposed by [Guo, et al. 2005b] showed how to utilize this type of dyads without introducing a biased result. After implementation of

the EM-Algorithm, $M_1^{1,*}$ is separated into $M_{1-1}^{1,*}$ (heterozygous parents who transmitted the B_1 allele but not the B_2 alleles) and $M_{1-2}^{1,*}$ (heterozygous parents who transmitted the B_2 allele but not the B_1 alleles). Note that $M_{1-1}^{1,*} + M_{1-2}^{1,*} = M_1^{1,*}$. Details of the EM procedure are available [Guo, et al. 2005b]. Consequently, the corresponding 2×2 table of the EM-HRR can then be jointly analyzed with unrelated cases and controls as discussed in the CHRR statistic.

BRESLOW-DAY TEST

Epstein et al. provided formal tests to determine whether trios and unrelated cases and controls can be combined. Therefore, two tests (1: unrelated cases can be combined with trios; 2: unrelated controls can be combined with trios) must be performed without any significant result to ensure validity of Epstein et al.'s strategy. Instead of using two likelihood ratio tests, we propose a single test to achieve the same objective.

We adopt the Breslow-Day test [Breslow and Day 1994] to test homogeneity of two odds ratios. The null hypothesis is that two odds ratios (one from trios and one from unrelated case-controls) are equivalent. Let $N_T = N_{11} + N_{13} + N_{21} + N_{23}$ denote total alleles from trios and $N_{CC} = N_{12} + N_{14} + N_{22} + N_{24}$ denote total allele from case-controls, then

$OR_{MH} = \frac{N_{11} \times N_{23} / N_T + N_{12} \times N_{24} / N_{CC}}{N_{13} \times N_{21} / N_T + N_{14} \times N_{22} / N_{CC}}$ and the Breslow-Day statistic is computed as $Q_{BD} = \frac{(N_{11} - E(N_{11}|OR_{MH}))^2 + (N_{12} - E(N_{12}|OR_{MH}))^2}{\text{var}(N_{12}|OR_{MH})}$, where E and var denote expected value and variance, respectively. When the null hypothesis is true, the statistic has an asymptotic chi-square distribution with 1 degree of freedom.

SIMULATIONS

Let "a" and "A" denote the disease and normal allele. Let D denote that an individual is diseased. Let f denote the probability of being affected when an individual carries 0 risk alleles (the phenocopy rate), and let K denote the genotype relative risk (GRR). For a recessive disease model, the penetrance functions are $P(D|AA) = P(D|Aa) = f$ and $P(D|aa) = K \times f$, where $0 \leq f \leq 1$ and $0 \leq K \times f \leq 1$. The disease prevalence is determined by these probabilities and the risk allele frequency. Similarly, for a dominant disease model, $P(D|AA) = f$ and $P(D|Aa) = P(D|aa) = K \times f$. For an additive disease model, $P(D|AA) = f$; $P(D|Aa) = K \times f$; $P(D|aa) = 2 \times K \times f$ or 1 if $K \times f > 0.5$.

We first considered additive, recessive, and dominant disease models in a homogeneous population. We simulated a general population with family trios and unrelated subjects according to the disease allele frequency (q), marker allele frequency of the B_1 allele (r), disequilibrium coefficient (δ), recombination fraction (θ), and penetrance functions. Among those trios, each father and mother was randomly assigned a missing status with probability p_f and p_m . So the expected number of complete trios, dyads, and monads ascertained from the simulated population are $100 \times (1-p_f) \times (1-p_m)$, $100 \times [(1-p_f) \times p_m + p_f \times (1-p_m)]$, and $100 \times p_f \times p_m$, respectively. Different probabilities for missing parents yielded similar results and the results are not provided.

After the homogeneous population was generated, 100 nuclear families with one affected offspring, 100 cases, and 100 controls were randomly selected from the simulated data. In an attempt to examine the admixture effect, we first simulated two populations with different disease and marker allele frequencies and then randomly ascertain 100 trios and sampled 100 cases and 100 controls from the admixed population. In Table 4, the disease and marker

allele frequency are both 0.3 (0.1) in “homo 1” (homo 2). For “admix 1” (admix 2), disease and marker allele frequencies are both 0.3 (0.2) for the first population and 0.5 (0.6) for the second population.

In Tables 4-6, the “TDT” column reports results using the traditional TDT test on the subset of complete trios only. The column headed “1-TDT” uses both the complete trios and one parent-offspring pairs. The “EM-HRR” column uses all three types of family data, which are complete trios, dyads and monads. The column marked “CC” uses unrelated case-control data only.

The column marked “SCOUT” represents the [Epstein, et al. 2005] method combining trios, dyads, unrelated cases, and unrelated controls. It is a test of association between SNP and disease provided that one failed to reject two null assumptions: 1) whether controls can be combined with trios 2) whether the cases can be safely combined with the trios and controls. Those replicates that rejected one or two of the null hypothesis were considered as null results to insure the validity of their strategy (See SCOUT software manual, page 11, for more details). Regardless of the true disease model in the simulations, the additive disease model was adopted in SCOUT analysis. This approach is commonly adapted for likelihood ratio tests for association when the underlying disease model is unknown. The “CHRR” column utilizes the combined data used in “EM-HRR” and “CC”. The column marked “CHRR_{BD}” excludes significant results of the Breslow-Day Test for the CHRR method, i.e. the replicates with significant p-values from the BD test are considered as null regardless of the significance of CHRR, which is analogous to the strategy of SCOUT.

We repeated the simulation 10,000 and 1,000 times to examine type-I error and power, respectively. Under the null hypothesis of no association ($\delta = 0$) in the presence of linkage ($\theta < 0.5$), the fraction of times that each test statistic exceeds the critical value, defined by the asymptotic distribution of the statistic, is the simulated type I error. The simulated power of each test is the proportion of test statistics in the total number of simulations exceeding the critical value under the alternative hypothesis.

RESULTS

In Table 4, we display the Type-I errors of the TDT, 1-TDT, EM-HRR, CC, SCOUT, CHRR, and CHRR_{BD} assuming no linkage ($\theta = 0.5$) and no association ($\delta = 0$). All methods have expected type-I error at approximately the nominal level of $\alpha = 0.05$ in a homogeneous population. Even when disease model was not correctly specified in SCOUT simulation, inflated type-I error was not observed for SCOUT. Therefore, both SCOUT and CHRR are valid tests for association combining case-parent trios and unrelated case-controls when population stratification is absent.

When admixture is present, TDT, 1-TDT, and EM-HRR remained valid tests. However, CC, SCOUT, CHRR, CHRR_{BD} had inflated Type-I errors with respect to degree of admixture. CHRR_{BD} noticeably reduced the Type-I errors compared to CHRR. In the simulations, CHRR_{BD} had a better protection against admixture compared to SCOUT in a more admixed population.

In Table 5 (Table 6), we display power for detecting linkage and association for common (rare) diseases in a homogenous population. Different recombination fractions θ , yielded similar results and the results are displayed in the Appendix, Table A.2. The TDT yielded the lowest power due to the exclusion of dyads and monads. The 1-TDT that utilizes both trios and dyads has greater power than the TDT. The EM-HRR, which incorporates trios, dyads, and monads, is more powerful than the 1-TDT, but less powerful than the CC. Even when disease model was not correctly specified, SCOUT had decent power detecting

association. However, when disease prevalence is higher, the relative risks and odd ratio estimates from trios and case-controls may differ. As a result, SCOUT and CHRR_{BD} may be less powerful than tests using only case-controls and CHRR_{BD} is generally more powerful than SCOUT. In contrast, the CHRR is consistently the more powerful test regardless of the disease prevalence.

DISCUSSION

In some genetic association studies, samples contain both parental and unrelated controls. In this circumstance, better power can be achieved by combining the information from all subjects in the study. Instead of analyzing only trios using family-based association tests or only unrelated subjects using a case-control study design, [Nagelkerke, et al. 2004] proposed a method that implemented a likelihood ratio test to combine the two different types of data. Since their strategy used an approximate analysis that requires strong assumptions of Hardy-Weinberg equilibrium (HWE), random mating, and a multiplicative disease model, [Epstein, et al. 2005] proposed another likelihood-based approach assuming that the disease is rare in an effort to combine trios with unrelated cases and controls.

In this article, we proposed a general and powerful test for association in the presence of linkage, which jointly analyzes family trios, dyads, monads and unrelated cases and controls. Through theoretical proof, we show that affected offspring can be pooled with diseased cases and parental controls can be treated as unrelated controls without the assumption that the disease is rare, as long as the trios and unrelated cases and controls are randomly sampled from a homogeneous population. According to the simulations, we demonstrate that the CHRR does not yield inflated type-I errors by pooling trios and unrelated subjects into one 2×2 table in a homogeneous population. When admixture is present, the Breslow-Day test can noticeably reduce the type-I error due to spurious association using case-controls. The simulation results suggest that the CHRR is more powerful than SCOUT for detecting association in the presence of linkage in a homogeneous population.

The additional advantage of the CHRR is that it is an allele-based association test and does not require the rare disease assumption. The reason is that regardless of the disease prevalence, the HRR for trios and the OR for case-controls are expected to be 1 under the null hypothesis. Therefore, the CHRR remains a valid test for association even when the disease prevalence is higher. Under the alternative hypothesis, the HRR may not approximate the true relative risk (RR) and reduces the statistical power. Nevertheless, simulations suggest that the CHRR has a higher power than the SCOUT under such scenario. Therefore, the CHRR is less restricted than Epstein et al.'s strategy under the null hypothesis and is more powerful under the alternative hypothesis. In addition, given the 2×2 contingency table design, the CHRR can be easily implemented in the widely used logistic regression framework to adjust for covariates and examine potential gene-environment interactions. Even with a multi-allelic marker with K alleles, the CHRR can be extended intuitively and effortlessly to a $K \times 2$ contingency table and the above logistic regression model can be implemented.

Acknowledgments

We would like to acknowledge the support of the Genomics Program at Children's Hospital Boston. This work was supported by in part National Heart, Lung and Blood Institute's Framingham Heart Study (Contract No. N01-HC-25195). We thank the two anonymous reviews for their insightful comments that greatly improved the manuscript.

APPENDIX

From Table 1, once can calculate the following:

$$P(B_1B_1|\text{unrelated controls}) = (X_1^2 + 2X_1X_3) / (1 - q^2)$$

$$P(B_1B_2|\text{unrelated controls}) = (2X_1X_2 + 2X_1X_4 + 2X_2X_3) / (1 - q^2)$$

$$P(B_2B_2|\text{unrelated controls}) = (X_2^2 + 2X_2X_4) / (1 - q^2)$$

$$P(B_1B_1|\text{unrelated cases}) = X_3^2 / q^2$$

$$P(B_1B_2|\text{unrelated cases}) = 2X_3X_4 / q^2$$

$$P(B_2B_2|\text{unrelated cases}) = X_4^2 / q^2$$

Therefore,

$$\begin{aligned} P(B_1|\text{unrelated controls}) &= P(B_1B_1|\text{unrelated controls}) + P(B_1B_2|\text{unrelated controls}) / 2 \\ &= (X_1 + pX_3) / (1 - q^2) = r - \frac{\delta q}{1 - q^2} \end{aligned}$$

$$\begin{aligned} P(B_2|\text{unrelated controls}) &= P(B_2B_2|\text{unrelated controls}) + P(B_1B_2|\text{unrelated controls}) / 2 \\ &= (X_2 + pX_4) / (1 - q^2) = (1 - r) + \frac{\delta q}{1 - q^2} \end{aligned}$$

Similarly,

$$\begin{aligned} P(B_1|\text{unrelated cases}) &= P(B_1B_1|\text{unrelated cases}) + P(B_1B_2|\text{unrelated cases}) / 2 \\ &= r + \frac{\delta}{q} \end{aligned}$$

$$\begin{aligned} P(B_2|\text{unrelated cases}) &= P(B_2B_2|\text{unrelated cases}) + P(B_1B_2|\text{unrelated cases}) / 2 \\ &= (1 - r) - \frac{\delta}{q} \end{aligned}$$

Table A.1

Joint display of probability density functions of the HRR and unrelated case-controls under the dominant disease model

	Affected offspring	Unrelated cases	Parental Controls	Unrelated controls
B ₁	$r + \frac{(1-\theta)(1-q)\delta}{q(2-q)}$	$r + \frac{(1-q)\delta}{q(2-q)}$	$r + \frac{(1-q)\theta\delta}{q(2-q)}$	$r - \frac{\delta}{q}$
B ₂	$(1-r) - \frac{(1-\theta)(1-q)\delta}{q(2-q)}$	$(1-r) - \frac{(1-q)\delta}{q(2-q)}$	$(1-r) - \frac{(1-q)\theta\delta}{q(2-q)}$	$(1-r) + \frac{\delta}{q}$
Sum	1	1	1	1

Table A.2

Power (%) of significance level $\alpha=0.05$ for rare diseases assuming the recombination fraction $\theta = 0.0001$. 100 trios, 100 cases, and 100 controls based on 1000 repetitions. GRR=2.

Disease Model	Allele freq	Baseline prevalence f	Disease Prevalence %	TDT	1-TDT	EM-HRR	CC	SCOUT	CHRR
Dominant	(0.3,0.3)	0.05	7.6	8.6	12.1	13.5	18.7	27.2	28.7
Additive	(0.3,0.3)	0.05	8.5	19.0	26.3	31.6	42.2	58.6	65.2
Recessive	(0.3,0.3)	0.05	5.5	6.6	8.4	8.9	9.7	12.8	14.6
Dominant	(0.3,0.1)	0.05	7.6	12.8	15.5	17.2	23.6	36.6	39.1
Additive	(0.3,0.1)	0.05	8.5	27.1	34.8	40.3	50.3	68	77.1
Recessive	(0.3,0.1)	0.05	5.5	7.7	8.4	9.7	9.7	12.5	14.9
Dominant	(0.1,0.3)	0.05	6.0	13.7	15.5	19.1	24.3	33.2	36.9
Additive	(0.1,0.3)	0.05	6.1	15.2	19.0	22.3	31.8	42.2	48.5
Recessive	(0.1,0.3)	0.05	5.1	5.3	5.3	5.8	5.3	5	6.1
Dominant	(0.1,0.1)	0.05	6.0	21.2	28.0	28.7	40.7	56.8	61.2
Additive	(0.1,0.1)	0.05	6.1	26.1	33.5	38.0	49.8	67.4	74.2
Recessive	(0.1,0.1)	0.05	5.1	5.0	6.9	5.3	4.9	5.3	6.0
Dominant	(0.3,0.3)	0.01	1.5	10.9	12.1	15.3	16.5	26.9	26.9
Additive	(0.3,0.3)	0.01	1.7	19.6	28.6	30.9	37.8	54.5	62.8
Recessive	(0.3,0.3)	0.01	1.1	6.9	7.0	7.7	10.1	10.6	13.2
Dominant	(0.3,0.1)	0.01	1.5	12.2	16.8	17.2	21.0	33.9	34.0
Additive	(0.3,0.1)	0.01	1.7	27.5	37.1	41.2	46.5	63.6	72.2
Recessive	(0.3,0.1)	0.01	1.1	8.7	9.2	9.7	10.1	13.4	15.7
Dominant	(0.1,0.3)	0.01	1.2	11.9	16.2	18.6	21.4	31.1	33.8
Additive	(0.1,0.3)	0.01	1.2	14.6	18.5	21.9	27.7	40.5	46.8
Recessive	(0.1,0.3)	0.01	1.0	5.5	5.5	5.4	4.9	3.9	5.4
Dominant	(0.1,0.1)	0.01	1.2	18.3	26.5	26.9	36.7	53.1	55.5
Additive	(0.1,0.1)	0.01	1.2	25.0	34.3	37.6	49.0	63.2	72.5
Recessive	(0.1,0.1)	0.01	1.0	5.7	5.1	4.4	4.2	4.3	4.9

Note: Missing rates for fathers and mothers are 0.2 and 0.4, respectively. The two numbers in the “Allele freq” column are disease and marker allele frequency, respectively. Disequilibrium coefficient $\delta = 0.1$ and 0.07 for allele frequency of 0.3 and 0.1, respectively.

REFERENCES

- Allen AS, Rathouz PJ, Satten GA. Informative missingness in genetic association studies: case-parent designs. *Am J Hum Genet.* 2003; 72:671–680. [PubMed: 12592606]
- Breslow, NE.; Day, NE. *The Design and Analysis of Cohort Studies.* Vol. II. Oxford University Press; USA: 1994. *Statistical Methods in Cancer Research.*
- Chen YH. New approach to association testing in case-parent designs under informative parental missingness. *Genet Epidemiol.* 2004; 27:131–140. [PubMed: 15305329]
- Clayton D. A generalization of the transmission/disequilibrium test for uncertain haplotype transmission. *Am J Hum Genet.* 1999; 65:1170–1177. [PubMed: 10486336]
- Curtis DR, Sham PC. A note on the application of the transmission disequilibrium test when a parent is missing. *Am J Hum Genet.* 1995; 56:811–812. [PubMed: 7887437]
- Devlin B, Roeder K. Genomic control for association studies. *Biometrics.* 1999; 55:997–1004. [PubMed: 11315092]
- Epstein MP, Allen AS, Satten GA. A simple and improved correction for population stratification in case-control studies. *Am J Hum Genet.* 2007; 80:921–930. [PubMed: 17436246]
- Epstein MP, Veal CD, Trembath RC, Barker JN, Li C, Satten GA. Genetic association analysis using data from trios and unrelated subjects. *Am J Hum Genet.* 2005; 76:592–608. [PubMed: 15712104]
- Ewens WJ, Spielman RS. The transmission/disequilibrium test: history, subdivision and admixture. *Am J Hum Genet.* 1995; 57:455–464. [PubMed: 7668272]
- Falk CT, Rubinstein P. Haplotype relative risks: An easy reliable way to construct a proper control sample for risk calculations. *Ann Hum Genet.* 1987; 51:227–233. [PubMed: 3500674]
- Guo CY, Cui J, Cupples LA. Impact of Non-Ignorable Missingness on Genetic Tests of Linkage and/or Association Using Case-Parents Trios. *BMC Genetics.* 2005a; 6(Suppl 1):S90. [PubMed: 16451706]
- Guo CY, Destefano AL, Lunetta KL, Dupuis J, Cupples LA. Expectation Maximization Algorithm Based Haplotype Relative Risk (EM-HRR): Test of Linkage Disequilibrium Using Incomplete Case-Parents Trios. *Hum Hered.* 2005b; 59:125–135. [PubMed: 15867473]
- Knapp M, Seuchter SA, Baur MP. The haplotype-relative-risk (HRR) method for analysis of association in nuclear families. *Am J Hum Genet.* 1993; 52(6):1085–1093. [PubMed: 8503442]
- Martin ER, Kaplan NL. A Monte Carlo procedure for two-stage tests with correlated data. *Genet Epidemiol.* 2000; 18:48–62. [PubMed: 10603458]
- Nagelkerke NJ, Hoebee B, Teunis P, Kimman TG. Combining the transmission disequilibrium test and case-control methodology using generalized logistic regression. *Eur J Hum Genet.* 2004; 12:964–970. [PubMed: 15340361]
- Ott J. Statistical properties of the haplotype relative risk. *Genet Epidemiol.* 1989; 6:127–130. [PubMed: 2731704]
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 2006; 38:904–909. [PubMed: 16862161]
- Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. *Am J Hum Genet.* 2000; 67:170–181. [PubMed: 10827107]
- Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.* 1999; 65:220–228. [PubMed: 10364535]
- PurcellSNealeBTodd-BrownKThomasLFerreiraMARBenderDMallerJSklarPde BakkerPIWDalyMJ and others. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet* 2007; 81:559–575 [PubMed: 17701901]
- RubinsteinPWalker.MKraussnerJCarrierCCarpenterCDoberstenMJNotkinsALMarkEMNechemiasCHau sknechtRU and others. HLA antigens and islet cell antibodies in gestational diabetes 19813271275

- Sebastiani P, Abad MM, Alpargu G, Ramoni MF. Robust transmission/disequilibrium test for incomplete family genotypes. *Genetics*. 2004; 168:2329–2337. [PubMed: 15611196]
- Sokal, RR.; Rohlf, FJ. W.H. Freeman; San Francisco: 1969. Biometry; the principles and practice of statistics in biological research; p. 612
- Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin dependent diabetes mellitus. *Am J Hum Genet*. 1993; 52:506–516. [PubMed: 8447318]
- Terwilliger JD, Ott J. *Hum Hered*. 1992; 42(6):337–346. [PubMed: 1493912]
- Thomson G. Mapping disease genes: family-based association studies. *Am J Hum Genet*. 1995; 57:487–498. [PubMed: 7668275]
- VealCDCap onFallenMHHeathEKEvansJCJonesAPatelSBurdenDTillmanDBarkerJNWN and others. Family-based analysis using a dense single-nucleotide polymorphism-based map defines genetic variation at PSORS1, the major psoriasis-susceptibility locus. *Am J Hum Genet* 2002 71:545–564 [PubMed: 12148091]
- Weinberg CR. Allowing for missing parents in genetic studies of case-parents triads. *Am J Hum Genet*. 1999; 64:1186–1193. [PubMed: 10090904]

Table 1

Haplotype probabilities

	Markers			
Disease Locus		B_1	B_2	Allele Frequency
	A	$X_1=pr - \delta$	$X_2=ps + \delta$	p
	a	$X_3=qr + \delta$	$X_4=qs - \delta$	q
	Allele Frequency	r	s	Sum = 1

Table 2

Joint Distribution of Transmitted and Non-Transmitted Marker Alleles for One Parent (Ott, 1989)

		Non-transmitted		
		B ₁	B ₂	Total
Transmitted	B ₁	$(r+\delta/q)r$	$(r+\delta/q)(1-r)-\theta\delta/q$	$r+(1-\theta)\delta/q$
	B ₂	$(1-r-\delta/q)r+\theta\delta/q$	$(1-r-\delta/q)(1-r)$	$(1-r)-(1-\theta)\delta/q$
	Total	$r+\theta\delta/q$	$(1-r)-\theta\delta/q$	1

Table 3

Joint display of probability density functions of Falk & Rubinstein's HRR (1987) and unrelated cases-controls under the recessive disease model

	Affected offspring	Unrelated cases	Parental controls	Unrelated controls
B ₁	$r + \frac{(1-\theta)\delta}{q}$	$r + \frac{\delta}{q}$	$r + \frac{\theta\delta}{q}$	$r - \frac{\delta q}{1-q^2}$
B ₂	$(1-r) - \frac{(1-\theta)\delta}{q}$	$(1-r) - \frac{\delta}{q}$	$(1-r) - \frac{\theta\delta}{q}$	$(1-r) + \frac{\delta q}{1-q^2}$
Sum	1	1	1	1

Table 4

Type-I error (%) of significance level $\alpha=0.05$ for 100 trios, 100 cases, and 100 controls based on 3000 repetitions. Recombination fraction $\theta = 0.5$, disequilibrium coefficient $\delta = 0$. $GRR=2$.

Disease Model	Admixture	Baseline prevalence f	Disease Prevalence	TDT	1-TDT	EM-HRR	CC	SCOUT	CHRR	CHRR _{BD}
Dominant	Homo 1	0.3	0.453	5.0	5.0	5.6	5.5	4.7	5.4	5.1
Additive	Homo 1	0.3	0.507	4.7	4.7	5.4	4.8	4.6	4.7	4.5
Recessive	Homo 1	0.3	0.327	5.0	5.2	5.7	5.1	4.6	5.1	4.8
Dominant	Homo 1	0.1	0.151	5.1	5.2	5.6	4.5	4.6	5.2	5.0
Additive	Homo 1	0.1	0.169	4.7	5.0	5.3	4.6	4.3	4.8	4.5
Recessive	Homo 1	0.1	0.109	5.0	5.2	5.5	4.8	4.5	5.3	5.0
Dominant	Homo 2	0.3	0.357	4.7	5.4	4.8	5.2	5.3	5.3	5.0
Additive	Homo 2	0.3	0.363	5.1	5.2	4.9	5.2	5.0	4.9	4.6
Recessive	Homo 2	0.3	0.303	4.9	5.2	4.4	4.6	4.8	4.7	4.5
Dominant	Homo 2	0.1	0.119	5.1	5.3	4.6	4.8	5.0	5.0	4.7
Additive	Homo 2	0.1	0.121	4.9	5.3	4.6	5.0	5.1	4.7	4.4
Recessive	Homo 2	0.1	0.101	5.0	5.3	4.7	4.5	5.0	4.7	4.5
Dominant	Admix 1	0.3	0.489	4.7	4.8	5.0	6.1	4.8	5.5	5.2
Additive	Admix 1	0.3	0.591	5.4	5.1	5.4	5.4	6.6	7.3	6.7
Recessive	Admix 1	0.3	0.351	5.1	5.1	5.1	5.7	4.8	5.3	4.9
Dominant	Admix 1	0.1	0.163	4.9	4.9	5.1	5.4	4.9	5.3	5.1
Additive	Admix 1	0.1	0.197	4.8	4.9	4.9	7.3	5.3	5.9	5.6
Recessive	Admix 1	0.1	0.117	4.9	4.9	5.0	5.6	4.7	5.3	5.1
Dominant	Admix 2	0.3	0.48	4.7	4.9	4.2	23.2	13.6	14.6	11.8
Additive	Admix 2	0.3	0.6	5.0	5.2	4.5	60.1	26.8	39.0	23.2
Recessive	Admix 2	0.3	0.36	5.0	5.1	4.3	15.4	8.5	9.6	8.2
Dominant	Admix 2	0.1	0.16	5.4	5.2	4.4	12.7	7.8	8.7	7.7
Additive	Admix 2	0.1	0.2	4.7	4.9	4.3	30.4	16.1	18.2	13.9
Recessive	Admix 2	0.1	0.12	5.0	5.4	4.2	10.6	5.9	7.1	6.5

Note: Disease and marker allele frequency are both 0.3 (0.1) in homo 1 (2). In admix 1 (2), disease and marker allele frequencies are both 0.3 (0.2) for the first population and 0.5 (0.6) for the second population. Missing rates for fathers and mothers are 0.2 and 0.4, respectively.

Table 5

Power (%) of significance level $\alpha=0.05$ for common diseases. 100 trios, 100 cases, and 100 controls based on 1000 repetitions. $GRR=2$.

Disease Model	Allele freq	Baseline prevalence f	Disease Prevalence %	TDT	1-TDT	EM-HRR	C/C	SCOUT	CHRR	CHRR _{BD}
Dominant	(0.3,0.3)	0.3	45.3	10.6	14.0	16.5	44.1	44.4	50.7	46.1
Additive	(0.3,0.3)	0.3	50.7	15.3	23.3	25.7	79.4	63.3	85.7	69.9
Recessive	(0.3,0.3)	0.3	32.7	6.7	7.6	8.5	14.2	15.3	18.4	17.3
Dominant	(0.3,0.1)	0.3	45.3	14.2	15.3	15.8	62.4	59.1	70.5	61.6
Additive	(0.3,0.1)	0.3	50.7	21.5	28.1	30.4	94.5	59.0	94.9	68.2
Recessive	(0.3,0.1)	0.3	32.7	7.4	9.7	10.1	17.8	17.1	21.7	20.3
Dominant	(0.1,0.3)	0.3	35.7	11.4	15.7	18.7	41.9	46.9	52.3	47.8
Additive	(0.1,0.3)	0.3	36.3	14.7	19.2	21.8	54.7	60.7	66.7	61.9
Recessive	(0.1,0.3)	0.3	30.3	5.6	4.9	4.1	4.8	4.7	5.6	5.2
Dominant	(0.1,0.1)	0.3	35.7	18.2	25.9	28.4	76.6	69.4	85.3	74.8
Additive	(0.1,0.1)	0.3	36.3	23.2	33.3	34.9	84.0	71.8	91.3	76.1
Recessive	(0.1,0.1)	0.3	30.3	4.5	5.5	5.0	5.7	5.6	7.8	7.1
Dominant	(0.3,0.3)	0.1	15.1	8.7	11.0	13.9	23.1	30.6	33.0	31.5
Additive	(0.3,0.3)	0.1	16.9	20.5	29.4	33.8	49.8	62.6	72.2	67.0
Recessive	(0.3,0.3)	0.1	10.9	5.7	6.9	7.3	9.7	10.7	13.1	12.7
Dominant	(0.3,0.1)	0.1	15.1	15.3	18.5	20.4	27.2	40.8	42.9	40.5
Additive	(0.3,0.1)	0.1	16.9	26.6	36.9	40.2	61.3	72.4	81.1	75.9
Recessive	(0.3,0.1)	0.1	10.9	9.3	8.9	10.1	13.2	14.8	18.7	18.0
Dominant	(0.1,0.3)	0.1	11.9	10.9	15.8	17.4	23.3	33.3	35.4	34.2
Additive	(0.1,0.3)	0.1	12.1	16.2	20.3	22.6	34.2	45.7	51.3	48.3
Recessive	(0.1,0.3)	0.1	10.1	5.3	5.3	5.2	5.0	4.1	5.8	5.6
Dominant	(0.1,0.1)	0.1	11.9	21.2	28.3	30.4	45.1	60.3	67.0	64.0
Additive	(0.1,0.1)	0.1	12.1	27.2	35.1	38.7	79.8	71.7	79.8	75.5
Recessive	(0.1,0.1)	0.1	10.1	4.9	6.5	5.3	5.9	5.0	5.9	5.5

Note: Missing rates for fathers and mothers are 0.2 and 0.4, respectively. The two numbers in the "Allele freq" column are disease and marker allele frequency, respectively. Recombination fraction $\theta=0$, disequilibrium coefficient $\delta=0.1$ and 0.07 for allele frequency of 0.3 and 0.1, respectively.

Table 6

Power (%) of significance level $\alpha=0.05$ for rare diseases. 100 trios, 100 cases, and 100 controls based on 1000 repetitions. $GRR=2$.

Disease Model	Allele freq	Baseline prevalence f	Disease Prevalence %	TDT	1-TDT	EM-HRR	C/C	SCOUT	CHRR	CHRR _{BD}
Dominant	(0.3,0.3)	0.05	7.6	9.9	12.7	14.4	20.0	28.0	29.5	27.6
Additive	(0.3,0.3)	0.05	8.5	19.5	26.0	32.0	43.5	57.4	64.5	59.8
Recessive	(0.3,0.3)	0.05	5.5	8.2	7.1	8.6	8.8	9.7	12.0	10.9
Dominant	(0.3,0.1)	0.05	7.6	13.5	17.8	20.1	22.4	36.0	38.5	36.5
Additive	(0.3,0.1)	0.05	8.5	26.2	35.2	39.8	52.7	69.9	76.5	72.3
Recessive	(0.3,0.1)	0.05	5.5	8.0	8.3	8.4	11.0	11.8	15.5	14.6
Dominant	(0.1,0.3)	0.05	6.0	10.9	14.0	16.7	19.7	31.9	33.3	31.3
Additive	(0.1,0.3)	0.05	6.1	17.5	20.9	22.8	29.7	39.2	46.3	43.7
Recessive	(0.1,0.3)	0.05	5.1	3.7	5.2	5.3	5.3	4.7	6.1	5.8
Dominant	(0.1,0.1)	0.05	6.0	19.6	29.0	31.2	38.6	56.8	62.5	59.3
Additive	(0.1,0.1)	0.05	6.1	28.4	35.7	38.0	52.6	66.0	75.2	70.8
Recessive	(0.1,0.1)	0.05	5.1	5.9	5.7	5.2	5.8	5.1	5.8	5.3
Dominant	(0.3,0.3)	0.01	1.5	10.8	14.1	16.5	16.4	25.9	26.9	25.0
Additive	(0.3,0.3)	0.01	1.7	18.0	25.1	27.4	37.9	53.2	60.8	58.1
Recessive	(0.3,0.3)	0.01	1.1	7.1	5.9	6.6	8.8	10.1	12.8	12.2
Dominant	(0.3,0.1)	0.01	1.5	12.4	16.2	16.9	21.4	33.8	35.8	34.0
Additive	(0.3,0.1)	0.01	1.7	27.9	36.0	41.6	45.2	65.5	73.2	69.3
Recessive	(0.3,0.1)	0.01	1.1	9.2	10.1	9.5	10.6	12.6	17.2	16.5
Dominant	(0.1,0.3)	0.01	1.2	11.5	16.0	18.6	17.8	31.5	33.6	32.0
Additive	(0.1,0.3)	0.01	1.2	14.5	19.9	22.6	26.3	40.0	44.6	42.2
Recessive	(0.1,0.3)	0.01	1.0	6.3	5.2	6.4	5.5	4.5	6.1	6.0
Dominant	(0.1,0.1)	0.01	1.2	17.6	26.6	28.5	36.5	51.0	56.0	53.2
Additive	(0.1,0.1)	0.01	1.2	24.1	33.8	33.8	47.7	62.2	69.7	65.5
Recessive	(0.1,0.1)	0.01	1.0	4.5	5.2	4.7	6.6	4.3	5.0	4.4

Note: Missing rates for fathers and mothers are 0.2 and 0.4, respectively. The two numbers in the "Allele freq" column are disease and marker allele frequency, respectively. Recombination fraction $\theta=0$, disequilibrium coefficient $\delta=0.1$ and 0.07 for allele frequency of 0.3 and 0.1, respectively.