

**Technical Report for the reference: Horvath S, Xu X, Lake SL, Silverman EK, Weiss ST, Laird NM (2004) Family based tests for associating haplotypes with general phenotype data: application to asthma genetics. Genet Epidemiol, Vol 26, No 1, 61-69**

Family Based Tests for Associating Haplotypes  
with General Phenotype Data:  
Application to Asthma Genetics

Steve Horvath<sup>1,2</sup>, Xin Xu<sup>3</sup>, Stephen L. Lake<sup>4</sup>,  
Edwin K. Silverman<sup>4,5</sup>, Scott T. Weiss<sup>3,4</sup>, and Nan M. Laird<sup>6</sup>

<sup>1</sup> Departments of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA; <sup>2</sup> Department of Biostatistics, School of Public Health, UCLA; <sup>3</sup> Program for Population Genetics, Harvard School of Public Health, Boston, MA; <sup>4</sup> Channing Laboratory, Brigham and Women's Hospital, Harvard Medical School, Boston; <sup>5</sup> Division of Pulmonary and Critical Medicine, Department of Medicine, Brigham and Women's Hospital, Boston; <sup>6</sup> Department of Biostatistics, Harvard School of Public Health, Boston

Address for correspondence: Steve Horvath

Department of Human Genetics, Gonda Research Center  
David Geffen School of Medicine, UCLA  
695 Charles E. Young Drive South, Box 708822  
Los Angeles, CA 90095-7088, USA  
Tel: (310) 825-9299  
Fax: (810) 277-7453  
E-mail: shorvath@mednet.ucla.edu

Short running title: haplotype FBAT

## ABSTRACT

We provide a general purpose family-based testing strategy for associating disease phenotypes with haplotypes when phase may be ambiguous and parental genotype data may be missing. These tests for linkage and association can be used in candidate gene studies with tightly linked markers. Our proposed weighted conditional approach extends the method described in Rabinowitz and Laird [2000] to multiple markers. It is attractive because it provides haplotype tests for family-based studies which are efficient and robust to population admixture, phenotype distribution specification and ascertainment based on phenotypes. It can handle missing parental genotypes and/or missing phase in both offspring and parents. It yields either haplotype-specific (univariate) tests or multi-haplotype (global) tests. This extension has been implemented in the freely available software haplotype FBAT. We used the haplotype FBAT program to test for associations between asthma phenotypes and single nucleotide polymorphisms (SNPs) in the beta-2 adrenergic receptor gene. Whereas no single SNP showed significant association with asthma diagnosis or bronchodilator responsiveness (quantitative trait), a haplotype-based global test found a highly significant association with asthma diagnosis (p-value  $< 0.00005$ ) and the measure of bronchodilator responsiveness (p-value  $< 0.016$ ).

**Key words:** FBAT; candidate region; admixture; association; transmission; missing; unphased; multi-locus

## INTRODUCTION

The use of haplotypes in association testing is increasingly important, especially when the informativeness of individual markers is low or when there are multiple genetic markers within a candidate gene. This raises an important issue in the context of family based association tests: how to properly account for inferred haplotypes, or haplotypes whose phase cannot be inferred with certainty? We briefly review four approaches for testing haplotypes with missing phase. The first approach uses one of several ‘haplotype reconstruction’ algorithms [Stephens et al., 2001] to derive a unique pair of haplotypes for each person in the sample. Although there are many haplotype reconstruction algorithms available, they are most suitable for homogeneous populations and can introduce bias in the presence of population admixture. Most do not exploit information on family structure, and do not give probabilistic assignments to other possible haplotypes. Tests based on this strategy are clearly subject to some error and methods for taking the potential misclassification into account are not available.

A second approach to testing is the basis of the TRANSMIT program [Clayton, 1999], which is further discussed below. A similar approach for case-control studies can accommodate quantitative endpoints [Schaid et al., 2002]. These approaches use maximum likelihood to estimate haplotype frequencies from the sample. Zhao et al. [2000] utilize phase unknown subjects in an extended TDT-type analysis, but assign probabilities to uncertain genotypes using a set of arbitrary weights in such a way that the resulting 2-way table of transmitted versus non-transmitted haplotypes is unbiased under the null hypothesis of no association. However, the approach by Zhao et al. [2000] does not allow missing parents.

A third approach uses conditioning arguments. For complete parental genotype data and phased genotype data, the conditioning principle as applied to family-based tests is described in Lazzeroni and Lange [1998]. Dudbridge et al. [2000] propose a test with both parents’ genotypes known that is based on a conditioning approach similar to one developed for a single marker in the general case of missing parents [Knapp, 1999; Rabinowitz and Laird, 2000]. As described below, the conditioning approach deals with missing parental information by conditioning on some function of the observed family genotype data, and computing the conditional distribution of the offspring genotypes. Here we present an approach, which builds on approach three and the weighting approach [Zhao et al., 2000]. When there is only one marker our approach reduces to the FBAT approach [Rabinowitz and Laird, 2000; Laird et al., 2000; Horvath et al., 2001]. In contrast to previous work [Dudbridge et al., 2000; Zhao et al., 2000; Knapp, 2001], we also describe how to deal with missing parental genotypes and with quantitative traits.

A fourth approach [Rabinowitz, 2002; Rabinowitz, 2003] differs in a fundamental way from approaches that characterize the null hypothesis in terms of conditional distributions: it uses the weaker condition that family-specific contributions to an asymptotically normal statistic should have expectation zero under the null hypothesis, regardless of the missing data. This may lead to increased power. But the cost of increased power can be the loss of known conditional distributions for the test statistic and thus exact conditional p-values or variance calculations are not available.

It is beyond the scope of this paper to systematically compare the alternative approaches. However we contrast our approach to TRANSMIT on some simulated and real data.

## STATISTICAL METHODS

### NOTATION

We consider genotype data that consist of several tightly linked genetic markers, i.e. we assume that there is no recombination between the markers. The markers can be bi- or multi-allelic. We use a comma to separate the alleles of an individual *haplotype*, e.g. a 3 marker haplotype is given by (1,2,2). When phase information is available, we separate the haplotypes by /, e.g. (1,2,2)/(2,1,2) denotes a phased 3 marker genotype. When phase information is unavailable, the genotypes of multiple markers are separated by a comma, e.g. an unphased 3 marker genotype is given by (1/2,1/2,2/2). If only one marker genotype is heterozygous in an unphased genotype, e.g., if  $g = (1/1, 1/2)$ , the phase assignment is trivial but in general, with data on relatives available, phase assignment is quite involved. Values of individual genotypes are labelled by lower case letters, e.g.  $g = (1/2, 1/2)$ . To stress that the phase is known, we use upper case letters, e.g.  $G = (1, 2)/(2, 1)$ . Our approach holds for multi-locus, multi-allelic markers, but we will use examples that involve only two di-allelic markers.

We use bold-face to denote vectors  $\mathbf{g}$  of *offspring* genotypes. Thus the number of components of the vector  $\mathbf{g}$  equals the number of offspring in the family under consideration. The values of  $\mathbf{g}$  will be used to denote points in the sample space of the conditional distribution of offspring genotypes. When the genotypes of each offspring are phased, we use the upper case notation  $\mathbf{G}$ .

Mating types are generally denoted by  $m$ . As with genotypes, when both parental mating types are phased, we will use upper case notation  $M$ . For example, an unphased and phased mating type is given by  $m = (1/2, 1/2) \times (1/2, 1/2)$  and  $M = (1, 1)/(2, 2) \times (1, 2)/(2, 1)$ , respectively. One of the aims of this paper is to study the case when the paternal, maternal or both parental genotypes may be missing. When both parental genotypes are missing we write  $m = \textit{unknown} \times \textit{unknown}$ , etc. For a vector of offspring genotypes  $\mathbf{g}$  and an unphased parental mating type  $m$ , we define  $CM(m|\mathbf{g})$  to be the set of phased parental *mating* types that are compatible with the vector of offspring genotypes  $\mathbf{g}$  and the mating type  $m$ .

**Example A:** Consider a family where two di-allelic markers have been genotyped in the father and his two offspring. The observed mating type is given by  $m = (1/1, 1/2) \times \textit{unknown}$ . Assume that the observed vector of offspring genotypes is given by  $\mathbf{g}^{\text{obs}} = ((1/1, 1/2), (1/2, 1/2))$ . It is clear that the phased genotype of the father and the first offspring is (1,1)/(1,2). We find that  $CM(m|\mathbf{g}^{\text{obs}}) = \{M_1, M_2, M_3, M_4\}$  where  $M_1 = (1, 1)/(1, 2) \times (1, 1)/(2, 1)$ ,  $M_2 = (1, 1)/(1, 2) \times (1, 1)/(2, 2)$ ,  $M_3 = (1, 1)/(1, 2) \times (1, 2)/(2, 1)$ ,  $M_4 = (1, 1)/(1, 2) \times (1, 2)/(2, 2)$ .

## CONDITIONING APPROACH

The general principle is to evaluate the distribution of test statistics using the conditional distribution of offspring genotypes under the null hypothesis, where the conditioning is on the sufficient statistic for any nuisance parameters in the model [Rabinowitz and Laird, 2000]. The potential nuisance parameters for nuclear families include the distribution of the phenotypes, the parental allele frequencies, and the model for ascertainment. By conditioning the offspring genotype distribution on the phenotypes, one eliminates sensitivity of the tests to misspecification of the phenotype distribution, and to ascertainment conditions which depend on the phenotypes. Conditioning on the parental genotypes eliminates sensitivity to population admixture; when parents' genotypes are unknown, we condition on the sufficient statistic for the parental allele frequencies, which are basically the set of observed genotypes in the offspring plus any observed parental genotypes. The conditioning approach underlies the FBAT software [Laird et al., 2000; Horvath et al., 2001], which can be used for phased genotype data by treating haplotypes as alleles of a multi-allelic marker. Here we extend the conditioning approach to tightly linked markers by conditioning on the sufficient statistic for resolving phase as well, which will again be based on observed parental and offspring genotypes. For simplicity of exposition, we restrict ourselves to the setting of possibly incomplete nuclear family data for testing  $H_0$ : no linkage and no association.

The following derivation of the conditional distribution is analogous to the one presented by Rabinowitz and Laird [2000] for a single marker. One can argue that when the founder genotype data are complete,  $x = \textit{the phased parental genotypes and traits}$  is the sufficient statistics for the nuisance parameters under the null hypothesis of no linkage and no association. But when the founder genotypes are *incomplete*, the minimal sufficient statistic is a function of the outcome  $y = \textit{observed offspring and parental genotypes and traits}$ . One can show that  $y$  and  $y'$  have the same value of the minimal sufficient statistic if and only if, for any value of the full data minimal sufficient statistic,  $x$ , either  $P(y|x)$  and  $P(y'|x)$ , are both equal to zero, or, the ratio  $P(y|x)/P(y'|x)$  is invariant to the choice of  $x$ . This formulation leads directly to steps 1-4 of the conditioning algorithm presented below. More specifically steps 1-4 yield an allowable set of those vectors of offspring genotypes  $\mathbf{g}_h$  that in conjunction with the observed traits have the same value of the minimal sufficient statistic as the observed vector of offspring genotypes  $\mathbf{g}^{\text{obs}}$ . Step 5 simply describes how to compute the conditional distribution  $P_{\text{cond}}(\mathbf{g}_h)$  given the observed data minimal sufficient statistic.

Example A (continued): After completing the steps of the algorithm described in the next section we find that the conditioning approach yields two allowable vectors of offspring genotypes  $\mathbf{g}_1 = ((1, 1)/(1, 2), (1/2, 1/2))$  and  $\mathbf{g}_2 = ((1/2, 1/2), (1, 1)/(1, 2))$ . Notice that the vectors of allowable offspring genotypes are comprised of genotypes that may or may not have phase information available. The conditional probability of these vectors of offspring genotypes is given by  $P_{\text{cond}}(\mathbf{g}_1) = P_{\text{cond}}(\mathbf{g}_2) = 0.5$ .

## THE FIVE STEPS OF THE ALGORITHM

This section provides the algorithmic steps for computing the conditional distribution of the vectors of offspring genotypes. The algorithm can deal with any number of markers with any number of alleles and any configuration of missing genotype data. It can only deal with zero-recombinant haplotypes. We will illustrate the steps of the algorithm by applying it to the following example.

**Example B:** Assume family data where two di-allelic markers have been genotyped in the father and his three offspring. The observed mating type is given by  $m = (1/1, 2/2) \times \text{unknown}$ . Assume that the vector of observed offspring genotypes is  $\mathbf{g}^{\text{obs}} = ((1/1, 1/1), (1/2, 1/2), (1/2, 1/2))$ .

**Step 1:** *Find all phased compatible mating types.* Denoting the vector of observed offspring genotypes by  $\mathbf{g}^{\text{obs}}$  and the observed parental mating type by  $m$ , step 1 can be written in terms of the CM function as  $CM(m|\mathbf{g}^{\text{obs}}) = \{M_1, M_2, \dots\}$  where  $M_k$  denotes the  $k$ -th compatible *phased* mating type.

Example B (continued): Since the first child is homozygous in the haplotype (1,1) it is clear that the phased genotype of the father must be (1, 1)/(2, 2) and that the maternal genotype contains the haplotype (1,1) as well. The genotype data of offspring 2 and 3 do not add any additional restrictions and we find  $CM(m|\mathbf{g}^{\text{obs}}) = \{M_1, M_2, M_3, M_4\}$  where  $M_1 = (1, 1)/(2, 2) \times (1, 1)/(1, 1)$ ,  $M_2 = (1, 1)/(2, 2) \times (1, 1)/(1, 2)$ ,  $M_3 = (1, 1)/(2, 2) \times (1, 1)/(2, 1)$ ,  $M_4 = (1, 1)/(2, 2) \times (1, 1)/(2, 2)$ .

**Step 2a:** *Find the minimal set of offspring genotypes which is consistent with all phased compatible mating types.* For each mating type  $M_k$  in  $CM(m|\mathbf{g}^{\text{obs}})$  find the sets of possible offspring genotypes  $\gamma_k$ . Take the intersection of the corresponding sets  $\gamma = \bigcap \gamma_k$ . If there is an ambiguous genotype in  $\gamma$ , check the phased mating types in  $CM(m|\mathbf{g}^{\text{obs}})$  to determine the phase. If the phase is the same for all mating types, set the ambiguous genotype equal to its corresponding phased value, otherwise leave it ambiguous. Note that  $\gamma$  contains at most 4 genotypes because each  $\gamma_k$  contains at most 4 genotypes. We denote the elements of  $\gamma$  by  $g^1, g^2$ , etc. For the  $k$ th mating type and the  $j$ th genotype  $g^j$  in  $\gamma$  define  $p_{jk} = P(g^j|M_k)$  to be the Mendelian transmission probability which is calculated elsewhere. These probabilities will be used in step 3.

Example B (continued): Set  $g^1 = (1, 1)/(1, 1)$  and  $g^2 = (1, 1)/(2, 2)$ . One can show that mating types  $M_1, \dots, M_4$  lead to  $\gamma_1 = \{g^1, g^2\}$ ,  $\gamma_2 = \{g^1, g^2, (1, 1)/(1, 2), (1, 2)/(2, 2)\}$ ,  $\gamma_3 = \{g^1, g^2, (1, 1)/(1, 2), (1, 2)/(2, 2)\}$ , and  $\gamma_4 = \{g^1, g^2, (2, 2)/(2, 2)\}$ . We find  $\gamma = \{g^1, g^2\}$ . The matrix  $(p_{jk})$  of relevant transmission probabilities is given by

$$P = \begin{pmatrix} 0.5 & 0.25 & 0.25 & 0.25 \\ 0.5 & 0.25 & 0.25 & 0.5 \end{pmatrix}$$

where the rows correspond to elements of  $\gamma$  ( $g^1, g^2$ ) and the columns correspond to the mating types  $M_1, \dots, M_4$ .

**Step 2b:** *Find the set of allowable vectors of offspring genotypes.* By using the genotypes defined in  $\gamma$ , form the set (referred to as list)  $L_1^*$  of all possible vectors of offspring genotypes.  $L_1$  is obtained from  $L_1^*$  by removing each vector of offspring genotypes  $\mathbf{g}$  for

which  $CM(m|\mathbf{g}) \neq CM(m|\mathbf{g}^{\text{obs}})$ . That is, remove any vector of offspring genotypes that leads to different inferences about possible parental genotypes than do the observed. This ensures all vectors of genotypes are consistent with the observed sufficient statistic for parental genotypes.

Example B (continued):  $L_1^*$  contains the following vectors of genotypes  $\{g^1\}$ ,  $\{g^2\}$ , and  $\{g^1, g^2\}$ . For the 3 offspring in the example, the notation  $\{g^1\}$  translates to  $(g^1, g^1, g^1)$  and  $\{g^1, g^2\}$  translates into three genotypes with at least one  $g^1$  and one  $g^2$ . At this point, distinguishing the order and number of offspring with a particular genotype is unimportant. Since  $CM(m|\{g^2\})$  contains the mating type  $M_5 = (1, 1)/(2, 2) \times (2, 2)/(2, 2)$  that is not contained in  $CM(m|\mathbf{g}^{\text{obs}})$ , we do not include  $\{g^2\}$  in  $L_1$ . Since  $CM(m|\{g^1\}) = CM(m|\{g^1, g^2\}) = CM(m|\mathbf{g}^{\text{obs}})$ , we find that  $L_1$  contains  $\{g^1\}$  and  $\{g^1, g^2\}$ .

**Step 3:** *Compute the conditional offspring genotype probability given the parental mating type.* Make a list  $L_2^*$  of possibly unphased offspring genotypes allowed by  $L_1$ . Since the following calculations will be invariant with respect to permuting the offspring genotypes, there is only a need to consider unordered lists of genotypes in  $L_2^*$ . For the  $h$ -th vector of genotypes and the  $k$ -th mating type define the matrix elements  $A_{hk}^* = P(h\text{-th vectors of genotypes}|M_k)$ . Note that  $A_{hk}^*$  is a product of the  $p_{jk}$ 's defined in step 2 since offspring genotypes are independent given the parental genotypes. Order the rows such that the observed vector of offspring genotypes  $\mathbf{g}^{\text{obs}}$  corresponds to the first row. Then define  $A_{hk} = A_{hk}^*/A_{1k}^*$ .

Example B (continued): The matrix  $A^*$  is given in table I. Taking ratios, we find that  $A_{hk}$  is given by

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0.25/.5 \\ 1 & 1 & 1 & 0.25^2/.5^2 \end{pmatrix}$$

where the rows correspond to the vectors of offspring genotypes in  $L_2^*$  and the columns to possible mating types, see table I.

**Step 4:** *Consider only vectors of genotypes in  $L_2^*$  whose corresponding row in the matrix  $A$  is constant.* Dropping non-constant vectors ensures that all the outcomes in this conditional distribution have the same probability for any possible mating type. This is a basic requirement for conditioning on a sufficient statistic. Denote the resulting list of unordered vectors by  $L_2$  and attach a prime ( $\prime$ ) to its indices. In step 5, the set of allowable vectors of offspring genotypes results from permuting the vectors of genotypes in  $L_2$ . For the  $h$ 'th unordered vector in  $L_2$ , the number of permutations ( $perm_{h\prime}$ ) can be computed as multinomial coefficient.

Example B (continued): Note that only the observed vector  $(g^1, g^2, g^2)$  corresponds to a constant row in the matrix  $A$ . Thus  $L_2$  contains only  $(g^1, g^2, g^2)$  and  $perm_1 = 3\text{-choose-}1 = 3$ .

**Step 5:** *Compute the conditional probabilities of each vector of offspring genotypes.* The set of allowable vectors of offspring genotypes is arrived at by permuting the vector of offspring genotypes in list  $L_2$ . The conditional probability of a permuted (ordered) vector of genotypes corresponding to the  $h$ 'th vector in  $L_2$  is  $P_{cond}(\mathbf{g}_{h\prime}) = A_{h\prime 1}^*/\sum_{m\prime} perm_{m\prime} A_{m\prime 1}^*$ . Here we defined  $P_{cond}$  with respect to the first column of  $A^*$  but the definition does not change when using another column (mating type).

Example B (continued): The ordered vectors of offspring genotypes in  $L_2$  are  $\mathbf{g}_1 = (g^1, g^2, g^2)$ ,  $\mathbf{g}_2 = (g^2, g^1, g^2)$ ,  $\mathbf{g}_3 = (g^2, g^2, g^1)$ . The conditional probability of the first vector is  $P_{cond}(\mathbf{g}_1) = .5^3/(3(.5)^3) = 1/3$ . Similarly one finds that  $P_{cond}(\mathbf{g}_2) = P_{cond}(\mathbf{g}_3) = 1/3$ .

## TEST STATISTICS AND WEIGHTS

Assume that there are  $N$  nuclear families indexed by  $i$ , each having  $n_i$  offspring, indexed by  $j = 1, \dots, n_i$ . For the  $i$ th family, the haplotype FBAT method determines the distribution of a possibly vector valued function  $S_i$  of the offspring genotypes and phenotypes  $S_i = S(\mathbf{g}_i, Y_i)$  where  $Y_i$  is a vector of the offspring traits and  $\mathbf{g}_i$  denotes an element of the allowable vectors of offspring genotypes found by the conditioning approach. Since we condition on the traits  $Y_i$ , the distribution of  $S_i$  depends on the distribution  $P_{cond}$  described above. Using the distribution of the  $S_i$ , one can define the score  $U = \sum_i \{S_i - E(S_i)\}$  and the variance  $V = \sum_i Var(S_i)$ . Note that the expected value and the variance are computed under the null hypothesis of no linkage and no association between the marker genotypes and any trait influencing gene. We define a Mantel-Haenszel type test statistic as  $U^T V^{-1} U$ , which asymptotically has a central  $\chi^2$  distribution with degrees of freedom equal to the rank of  $V$ . When dealing with a univariate marker coding, one can also define a Z-statistic  $Z = U/\sqrt{V}$ . Under the null hypothesis,  $Z$  will have an asymptotic standard normal distribution.

We will proceed with a form of  $S_i$  that is frequently used [Clayton, 1999; Whittemore and Tu, 2000; Schaid et al., 2002]. For simplicity, we suppress the index  $i$  in the following. In the case of a *phased* vector of offspring genotypes  $\mathbf{G} = (G^1, G^2, \dots)$ , we define

$$S = \sum_j Y_j X(G^j)$$

where  $Y_j$  is the trait,  $X(G^j)$  is a general encoding of the phased genotype value of the  $j$ th offspring, and summation of over the  $n_i$  offspring. Here  $Y_j$  can be a dichotomous trait with values 0 and 1 or a suitably defined quantitative trait. The genotype coding  $X(G^j)$  can denote any univariate function of (phased) genotypes, or it can accommodate multiple haplotypes by turning  $X(G^j)$  into a vector as discussed below.

Now consider the case of a vector of possibly unphased offspring genotypes  $\mathbf{g}$ . We propose to generalize the basic form of the test statistic to the case of a vector of unphased genotypes by using a weighted approach, which is similar to the population based approach. We will assign weights to all the possible phased genotypes which are consistent with any ambiguous genotype in  $\mathbf{g}$ . Let us be specific. For an individual *unphased* genotype  $g^j$  define the genotype coding  $X()$  by  $X(g^j) = \sum_k X(G^{jk}) w_{G^{jk}}$  where  $k$  sums over the set of possible phased genotypes  $G^{jk}$  that are compatible with  $g^j$ . The weights are constraint to satisfy  $\sum_k w_{G^{jk}} = 1$ . Importantly, note that we use the weights only to evaluate the test statistic and not to assign phase to the offspring or parental genotypes. This device is also used to calculate  $E(S_i)$  and  $Var(S_i)$ . Note that  $E(S - E(S)) = 0$  under the null hypothesis by construction. Unbiasedness follows because the distribution of  $S$  is computed conditional on the sufficient statistics for any nuisance parameters under the null hypothesis. The weights are estimated from the sufficient statistics, and not from disease status or offspring

transmissions, hence the statistics are unbiased conditionally for each set of possible weights, and thus overall. This unbiasedness is reinforced in the simulations.

The choice of weights will affect the power. It is natural to estimate the weights  $w_{G^{jk}}$  by the conditional probability of observing  $G^{jk}$  given that it is compatible with  $g^j$ . In our software, we use an expectation-maximization (EM) algorithm to estimate haplotype frequencies from which the weights are calculated. The EM algorithm maximizes the likelihood of the phased genotype frequencies, based on all the families observed genotypes, and computed under the null hypothesis. The resulting weights have been used in our simulation study and real data application.

Example A (not B, continued): We find that  $X((1/2, 1/2))$  is given by  $X((1, 1)/(2, 2))w + X((1, 2)/(2, 1))(1-w)$  where  $w$  is the weight associated with the phased genotype  $(1, 1)/(2, 2)$ . If  $X()$  counts the number of  $(1, 1)$  haplotypes then  $S(\mathbf{g}_1, Y) = Y_1 + Y_2w$  and  $S(\mathbf{g}_2, Y) = Y_1w + Y_2$ . Since  $P_{cond}(\mathbf{g}_1) = P_{cond}(\mathbf{g}_2) = .5$  one can easily verify that  $E(S) = (Y_1 + Y_2)(1 + w)/2$  and  $Var(S) = E(S^2) - E(S)^2$  where  $E(S^2) = (Y_1 + Y_2w)^2/2 + (Y_1w + Y_2)^2/2$ . Since the observed vector of offspring genotypes is  $\mathbf{g}_1$ , we find  $U = Y_1 + Y_2w - (Y_1 + Y_2)(1 + w)/2$ . If both offspring have the same trait value ( $Y_1 = Y_2$ ) then  $U = S - E(S) = 0$  and  $Var(S) = 0$  for any  $w$ , i.e., the family is not informative about the  $(1, 1)$  haplotype.

## EMPIRICAL VARIANCE ESTIMATOR

In some cases, as in the dissection of putatively linked regions with association methodology, a null hypothesis that allows for linkage but no association may be warranted [Martin et al., 2000; Lake et al., 2000]. When testing for association in the presence of linkage or when using a sample that consists of pedigree data, a robust variance estimator can be used to account for the correlations in marker genotypes that exist under the null hypothesis. When nuclear families are sampled the robust variance estimator is  $V_R = \sum_i \{S_i - E(S_i)\} \{S_i - E(S_i)\}'$ . There is a small technical challenge: in principle, one should compute the expected value with respect to a conditional distribution corresponding to the null hypothesis of no association but possibly linkage. To arrive at this distribution, one would have to alter the algorithm introduced in this paper. It turns out that the resulting expected values are identical to those that result from using the conditional distribution derived above for the case of the null hypothesis of no linkage and no association [Lake et al., 2000]. Thus there is no need to compute a new conditional distribution when using empirical variance estimators. In the case of extended pedigrees, the nuclear families embedded in a pedigree are used to compute the test statistic. The sum of these test statistics define the contribution from a specific pedigree. The associated robust variance estimator is defined as  $V_R = \sum_p [\sum_i \{S_{pi} - E(S_{pi})\} \{S_{pi} - E(S_{pi})\}']$  where  $p$  indexes pedigrees and  $i$  indexes nuclear families embedded in a pedigree.

## SIMULATION STUDY

The power and the robustness to population admixture of haplotype FBAT were explored via simulation for different study designs where a disease locus and marker loci are simulated [Boehnke and Langefeld, 1998; Knapp, 1999]. We assumed an additive model for the di-allelic disease locus. The disease prevalence  $K$  was fixed at 5% and the attributable fraction was set to 50%. For testing the validity, we simulated data under the null hypothesis of no linkage and no association. The eight haplotypes based on the possible combinations of three di-allelic markers were simulated. For the power simulations, the marker haplotypes (111), (112), (121), (122), (211), (212), (221), (222) were given population frequencies of 0.35, 0.20, 0.20, 0.10, 0.05, 0.05, 0.04 and 0.01, respectively, and the 111 haplotype was associated with the disease allele. The degree of association was parameterized with  $C$ , the frequency difference of the haplotype, comprised of the disease allele and the associated marker haplotype, between randomly selected cases and controls [Knapp, 1999]. To verify the robustness of haplotype FBAT to population admixture, a second population was sampled. The marker haplotype frequencies for the second population are the same as the frequencies described above, except that the frequencies for the (111) and (122) haplotypes are switched. The disease prevalence for the second population was 15%. Each simulated sample contained 600 affecteds from different family structures described in table II. Global tests of haplotype association and haplotype-specific tests of association for the (111) haplotype were performed on each simulated sample. For comparison purposes, the program TRANSMIT was also used to test for haplotype associations. Testing was performed with a type I error rate of 0.05. To study the validity and power of the tests we used 2500 and 500 data sets, respectively.

## RESULTS

The results of the simulations are presented in table II. Whereas haplotype FBAT is robust to population admixture, TRANSMIT is susceptible to type I error rate inflation when parental information is missing. For example, at the nominal 0.05 type I error rate, the empirical type I error rate for the haplotype-specific test is 0.05 for haplotype FBAT and is 0.082 for TRANSMIT when the sampled families consist of discordant sibpairs and one parent. The power simulations indicate that haplotype FBAT has power that is comparable to TRANSMIT when the data consist of family trios or discordant sibships with two affected and one unaffected offspring and no parents. When discordant sibships with completely or partially missing parental information are sampled, haplotype FBAT has reduced power compared to TRANSMIT, which may be due to 2 reasons: first, TRANSMIT may have a slightly inflated significance level in this setting; second, a number of families are uninformative for haplotype FBAT, as opposed to TRANSMIT which uses population haplotype frequencies to construct the test statistic.

From table II we gather that the haplotype specific test is more powerful than the global test. However the haplotype specific test did not adjust for the number of haplotypes considered. It is not surprising that the haplotype specific test would have greater power since we simulated only one high risk haplotype. Future studies should also consider genetic

models with two or more high-risk haplotypes.

## ASTHMA STUDY

We used haplotype FBAT to test for associations between asthma phenotypes and SNPs (-654, -47, +46, +79, +252, +491) in the Beta-2 Adrenergic Receptor gene ( $\beta_2AR$ ) with data from the Childhood Asthma Management Program [CAMP Research Group, 2000]. Asthma diagnosis and a measure of bronchodilator responsiveness (a quantitative trait coded as normal score) were tested in 652 nuclear families consisting of 2011 individuals. Polymorphisms in  $\beta_2AR$  have been associated with a number of asthma phenotypes [Tashkin et al. 1982], but association results have been inconsistent across different studies. In the CAMP study, eight SNPs in the  $\beta_2AR$  were genotyped, but only seven were used for testing purposes due to lack of informativeness in one polymorphism. The linkage disequilibrium as measured by Lewontin's  $D^2$  [Devlin and Risch, 1995] ranged from 0 to 0.83. A complete analysis of the  $\beta_2AR$  SNPs using previously available methods will be reported elsewhere [Silverman et al., 2003].

Associations between single SNPs and the asthma phenotypes were tested using the FBAT program [Laird et al., 2000; Horvath et al., 2001]. The smallest  $p$ -value (not adjusted for multiple comparisons) from the single SNP tests are reported for both the asthma diagnosis and the measure of bronchodilator responsiveness. Haplotypic associations were tested using haplotype FBAT and (for the asthma diagnosis phenotype) TRANSMIT. The measure of bronchodilator responsiveness was normalized for the tests. The number of informative families required for a haplotype to be included in the haplotype FBAT test statistic was set at 15. In TRANSMIT, the criteria for inclusion of haplotypes is based on the estimated haplotype frequencies. So, while an equivalent threshold for the two tests is not available, a haplotype frequency threshold of 0.13% in TRANSMIT produces a haplotype set for analysis that is similar to the haplotype set in haplotype FBAT. Table III contains the test results. Whereas no single SNP showed significant association between asthma diagnosis or the measure of bronchodilator responsiveness at the 0.05 significance level, the haplotype-based global tests of association for asthma diagnosis were highly significant (haplotype FBAT:  $\chi^2(9) = 35.62$ ,  $p < 0.00005$ , TRANSMIT:  $\chi^2(11) = 34.24$ ,  $p = 0.0003$ ). Note that the  $\chi^2$  statistic values of the two tests are very similar. The discrepancy between the  $p$ -values is due to different degrees of freedom, which in turn are determined by the different haplotype inclusion criteria mentioned above. The haplotype test for association between  $\beta_2AR$  and the measure of bronchodilator responsiveness was also significant ( $p < 0.016$ ). Since TRANSMIT cannot deal with quantitative traits at this point, the corresponding  $p$ -value is not available (NA). Table IV contains the haplotype patterns and frequencies of the 9 most informative haplotypes. For each haplotype it lists the Z-statistic and the corresponding  $p$ -value for the dichotomous (asthma diagnosis) and the quantitative (bronchodilator responsiveness) phenotype. A significant  $p$ -value and a positive (negative) Z-statistic are indicative of a high risk (protective) haplotype.

## DISCUSSION

Different approaches to handling missing parental data and/or phase in parents and offspring differ according to how the conditional distribution of  $X(g)$  is computed. The TRANSMIT software [Clayton, 1999] takes a population based approach based on likelihoods; estimates of the haplotype frequencies from the founder data are used to compute the probability of each possible phased parental genotype, conditional on the information observed in the family's genotypes. These conditional probabilities are used to attach weights to any unknown phased genotype in the family. Thus a very attractive feature of TRANSMIT is that data from all families can be used, even if, for example, one observes only one child's genotype and both parental genotypes are missing. We have shown that TRANSMIT can lead to a slightly elevated false positive rate in the presence of population admixture. While the bias may be small in most instances, the attractive feature of haplotype FBAT is that it is completely robust to population admixture so that this issue does not arise. We find that the TRANSMIT and haplotype FBAT have practically identical power when parental genotypes are available. Different from haplotype FBAT, TRANSMIT does not handle quantitative traits at this point.

The haplotype FBAT approach has no upper limit on the number of markers that can be analyzed but it is worth repeating that the markers should be tightly linked since the method assumes no recombination. If many markers are available, it may be more efficient to select only a subset of them. For example, one may use haplotype tagging SNPs (htSNPs), i.e., SNPs which capture the majority of the LD structure [Johnson et al., 2001; Stram et al., 2003]. Alternatively, markers could be chosen on the basis of their putative biological relevance.

As noted above, the test statistics involve computing  $X(G)$ , where  $G$  is a phased genotype. Since a haplotype can be viewed as an allele in a single multi-allelic marker all multi-allelic genotype codings, e.g. additive, recessive, dominant, can be used in our software. With haplotypes, more efficient strategies for testing many alleles are desirable, see for example [Clayton and Jones, 1998; Seltman et al., 2001; Hoh et al., 2001; Schaid, 2002; Cordell and Clayton, 2002]. Many of these approaches could be implemented quite straightforwardly by defining  $X()$  to be a haplotype grouping function.

## ELECTRONIC DATABASE INFORMATION

The haplotype FBAT method has been integrated into the FBAT program version 1.4. It is invoked with the `hbat` command. The program and its documentation can be downloaded from [www.biostat.harvard.edu/~fbat/default.html](http://www.biostat.harvard.edu/~fbat/default.html).

## ACKNOWLEDGEMENT

We thank all families for their enthusiastic participation in the Camp Genetics Ancillary Study, supported by the NHLBI grants N01-HR-16049. We also acknowledge the CAMP

investigators and research team, supported by NHLBI, for the collection of CAMP Genetics Ancillary Study data. Additional support for research came from grant SCOR P50 HL67664 from the NHLBI.

The research was supported in part by NIH grant 2R01MH059532-04A1.

## REFERENCES

- Boehnke M, Langefeld CD. 1998. Genetic association mapping based on discordant sib-pairs: the discordant alleles test (DAT). *Am J Hum Genet* 62:950-961.
- Clayton D. 1999. A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *Am J Hum Genet* 65:1170-1177.
- Clayton D and Jones H. 1999. Transmission/disequilibrium tests for extended marker haplotypes. *Am J Hum Genet* 65:1161-1169.
- Childhood Asthma Management (CAMP) Research Group. 2000. Long-term effects of budesonide or nedocromil in children with asthma. *N Engl J Med* 343(15):1054-1063.
- Cordell HJ, Clayton DG. 2002. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in Type 1 Diabetes. *Am J Hum Genet* 70:124-141.
- Devlin B, Risch N. 1995. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* 29(2):311-22.
- Dudbridge F, Koeleman BPC, Todd JA, Clayton DG. 2000. Unbiased application of the transmission/disequilibrium test to multilocus haplotypes. *Am J Hum Genet* 66:2009-2012.
- Hoh J, Wille A, Ott J. 2001. Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Res* 11:2115-2119.
- Horvath S, Xu X, Laird NM. 2001. The family based association test method: strategies for studying general genotype-phenotype associations. *Eur J Hum Genet* 9:301-306.
- Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, Twells RC, Payne F, Hughes W, Nutland S, Stevens H, Carr P, Tuomilehto-Wolf E, Tuomilehto J, Gough SC, Clayton DG, Todd JA. 2001. Haplotype tagging for the identification of common disease genes. *Nat Genet* 29(2):233-7.
- Stram DO, Haiman CA, Hirschhorn JN, Altshuler D, Kolonel LN, Henderson BE, Pike MC. 2003. Choosing Haplotype-Tagging SNPS Based on Unphased Genotype Data Using a Preliminary Sample of Unrelated Subjects with an Example from the Multiethnic Cohort Study. *Hum Hered* 55(1):27-36.

- Knapp M. 1999. The transmission/disequilibrium test and parental genotype reconstruction: the reconstruction-combined transmission/disequilibrium Test. *Am J Hum Genet* 64:861-870.
- Knapp M. 2001. A family-based test for association in the presence of linkage with multiple tightly linked markers in nuclear families with multiple affected children. Talk, 10-th Conf of the International Genet Epidemiol Soc (IGES), Garmisch-Partenkirchen, Germany.
- Laird NM, Horvath S, Xu X. 2000. Implementing a unified approach to family based tests of association. *Genet Epidemiol Suppl* 19:S36-S42.
- Lake S, Blacker D, Laird NM. 2000. Family based tests in the presence of association. *Am J Hum Gen* 67:1515-1525.
- Lazzeroni LC, Lange K. 1998. A conditional inference framework for extending the transmission/disequilibrium test. *Hum Hered* 48(2):67-81.
- Martin ER, Monks SA, Warren LL, Kaplan NL. 2000. A Test for Linkage and Association in General Pedigrees: The Pedigree Disequilibrium Test. *Am J Hum Genet* 67:146-154.
- Rabinowitz D. 2002. Adjusting for population heterogeneity and misspecified haplotype frequencies when testing non-parametric null hypotheses in statistical genetics. *J Am Stat Assoc* 97:742-751.
- Rabinowitz D. 2003. Adjusting for population heterogeneity: A framework for characterizing statistical information and developing efficient test statistics. *Genetic Epidemiol* 24(4):284-290.
- Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. 2002. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet* 70(2):425-34.
- Seltman H, Roeder K, Devlin B. 2001. Transmission/disequilibrium test meets measured haplotype analysis: family-based association analysis guided by evolution of haplotypes. *Am J Hum Genet* 68:1250-1263.
- Silverman EK, Kwiatkowski DJ, Sylvia JS, Lazarus R, Drazen JM, Lange C, Laird NM, Weiss ST. 2003. Family-based association analysis of Beta-2 adrenergic receptor polymorphisms in the Childhood Asthma Management Program. *J Allergy Clin Immunol* (In Press).
- Stephens M, Smith N, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 68:978-989.
- Tashkin DP, Conolly ME, Deutsch RI, Hui KK, Littner M, Scarpace P, Abrass I. 1982. Subsensitization of beta-adrenoceptors in airways and lymphocytes of healthy and asthmatic subjects. *Am Rev Respir Dis* 125(2):185-93.

Whittemore AS, Tu IP. 2000. Detecting disease genes using family data (I). Likelihood-based theory. *Am J Hum Genet* 66:1328-1340.

Zhao H, Zhang S, Merikangas KR, Trixler M, Wildenauer DB, Sun F, Kidd KK. 2000. Transmission/disequilibrium tests using multiple tightly linked markers. *Am J Hum Genet* 67:936-946.

TABLE I. Computing the matrix  $A^*$  (step 3 of the conditioning algorithm) for example B.

Vectors of Genotypes $L_2^*$	Mating Type			
	$M_1$	$M_2$	$M_3$	$M_4$
$(g^1, g^2, g^2)$	$.5^3$	$.25^3$	$.25^3$	$(.25)(.5)^2$
$(g^1, g^1, g^2)$	$.5^3$	$.25^3$	$.25^3$	$(.25)^2(.5)$
$(g^1, g^1, g^1)$	$.5^3$	$.25^3$	$.25^3$	$.25^3$

Table II. Empirical type I error rate and power of haplotype FBAT and TRANSMIT.  $C$  measures haplotype association. Haplotype-specific (hap) and global (global) test evaluated on family trios (trios), discordant sibpairs with both parents missing (AU0), discordant sibpairs with one parent missing (AU1), and discordant sibships with two affected and one unaffected (AAU0).

Approach	$C$	trio		AU0		AU1		AAU0	
		hap	global	hap	global	hap	global	hap	global
haplo FBAT	$0^a$	0.0512	0.0508	0.0508	0.0488	0.0500	0.0516	0.0506	0.0502
	.075	0.932	0.738	0.706	0.386	0.668	0.344	0.706	0.476
TRANSMIT	$0^a$	0.0548	0.0564	0.0764	0.0660	0.0820	0.0904	0.0917	0.0924
	0.075	0.936	0.754	0.846	0.516	0.718	0.356	0.716	0.380

$a$ : For these simulations, admixed samples were generated.

**TABLE III.** Global (multi-haplotype) tests for association between SNPs in the  $\beta_2AR$  gene and asthma-related phenotypes.

phenotype	FBAT <sup>a</sup>			haplotype FBAT			TRANSMIT		
	$\chi^2$	<i>df</i>	<i>p</i> -value	$\chi^2$	<i>df</i>	<i>p</i> -value	$\chi^2$	<i>df</i>	<i>p</i> -value
asthma diagnosis	2.20	1	0.138	35.62	9	< 0.00005	34.24	11	0.0003
bronch. response	3.39	1	0.066	20.33	9	0.016			NA

*a*: Single marker FBAT analysis. Only the most significant result has been reported.

**TABLE IV. Haplotype-specific univariate FBAT statistics (Z-statistics) for studying the relationship between asthma phenotypes and haplotypes that were present in 15 or more informative families (inf. fams).**

haplotype	freq.	inf. fams	Z-statistic (p-value)	
			asthma	bronch. resp.
ATACGCC	0.3574	295	0.54 (0.590)	-0.60 (0.55)
GCGGGCC	0.3549	271	0.63 (0.530)	-0.20 (0.84)
GTGCACA	0.1798	209	-0.75 (0.450)	-1.49 (0.14)
GTACGCC*	0.0357	48	2.46 (0.014)	1.30 (0.19)
GTGCACC	0.0293	48	0.41 (0.680)	1.93 (0.054)
GTGCATA	0.0118	18	-0.45 (0.650)	0.03 (0.98)
ATGCGCC	0.0063	30	-0.02 (0.980)	0.87 (0.38)
GTGCGCA*	0.0048	17	-2.65 (0.008)	2.09 (0.037)
GTACACA	0.0009	24	1.46 (0.140)	-1.21 (0.230)

\*: significant at level 0.05