

A Unified Approach to Adjusting Association Tests for Population Admixture with Arbitrary Pedigree Structure and Arbitrary Missing Marker Information

Daniel Rabinowitz^a Nan Laird^b^aDepartment of Statistics, Columbia University, New York, N.Y., and ^bDepartment of Biostatistics, Harvard School of Public Health, Boston, Mass., USA

Key Words

Haplotype relative risk · Identity by descent · Linkage analysis · Linkage disequilibrium · Transmission/disequilibrium

Abstract

A general approach to family-based examinations of association between marker alleles and traits is proposed. The approach is based on computing p values by comparing test statistics for association to their conditional distributions given the minimal sufficient statistic under the null hypothesis for the genetic model, sampling plan and population admixture. The approach can be applied with any test statistic, so any kind of phenotype and multi-allelic markers may be examined, and covariates may be included in analyses. By virtue of the conditioning, the approach results in correct type I error probabilities regardless of population admixture, the true genetic model and the sampling strategy. An algorithm for computing the conditional distributions is described, and the results of the algorithm for configurations of nuclear families are presented. The algorithm is applicable with all pedigree structures and all patterns of missing marker allele information.

Copyright © 2000 S. Karger AG, Basel

Introduction

With random mating in a closed population, association between unlinked loci is halved every generation. Association for linked loci, on the other hand, dissipates more slowly. This phenomenon is the basis for using association between marker alleles and traits to detect linkage. With nonrandom mating, migration or admixture of subpopulations, however, even when the marker is unlinked to any trait locus, association between marker alleles and alleles at trait loci is possible. The strategy of using association between marker alleles and traits can therefore result in spurious evidence of linkage.

A variety of methods have been developed for using family data to avoid spurious evidence of linkage when examining association between genotypes and traits. In most cases, these methods are valid not only regardless of population admixture, but also regardless of the genetic model and sampling plan. See, for example, Falk and Rubinstein [1987], Spielman et al. [1993] or Terwilliger and Ott [1992].

Generally, although these methods may be motivated by concepts such as ‘using the other parental allele as a control’ or as applications of well-known nonparametric procedures such as McNemar’s test, they may be viewed as implicitly involving conditioning on parental marker alleles and on traits, and examining whether traits are pre-

dictive of patterns of transmission of parental alleles to offspring. The conditioning approach is not immediately applicable, however, when parental marker data are not available. Curtis and Sham [1995] note that there are subtle difficulties involved with the strategy of inferring parental marker data from offspring data and then proceeding as if the parental data had been observed. Spielman and Ewens [1998] and Boehnke and Langefeld [1998] present approaches that may be used in nuclear families when parental marker information is not available, but the approaches are applicable only in the context of discordant sibships and dichotomous traits. Schaid and Li [1997] present a likelihood-based approach for the case of missing parental marker data, but their approach involves estimation of marker allele frequencies and thus can be subject to bias when assumptions underlying the estimation procedure (such as the existence of a homogeneous marker frequency) are violated. Here is presented a general approach to adjusting for population admixture. The approach is valid regardless of the genetic model and sampling plan, allows incorporation of covariates, and may be used with arbitrary phenotypes, arbitrary pedigree structure and arbitrary patterns of missing marker allele information.

The approach presented here is based on the fundamental statistical method of conditioning on sufficient statistics for the null hypothesis. Sufficient statistics for a set of models are defined by the property that the conditional distribution given the statistics are the same for all models in the set. Thus, if one computes *p* values conditionally given the sufficient statistics for the models in the null hypothesis, then, for all models in the null hypothesis, the computations are the same. That is, the *p* values result in rejecting the null hypothesis with the correct type I error rate regardless of which model in the null hypothesis is true.

The proposal here is to apply directly this fundamental statistical method. Given a test statistic that is sensitive to association between traits and marker alleles, the approach is to compute *p* values by comparing the test statistic to its conditional distribution given the minimal sufficient statistic under the null hypothesis for population admixture, the sampling plan and the genetic model. No restrictions are placed on the form of the test statistics so they may be chosen to reflect information about mode of inheritance, they may be applied with any kind of phenotype and they may be designed to incorporate covariate information. Through conditioning on a sufficient statistic, the approach is valid, in the sense of resulting in correct type I error rates regardless of patterns of population

admixture, the sampling plan, and the genetic model. See, for example, Cox and Hinkley [1974]. Because conditioning on the minimal sufficient statistics is generally applicable, the approach may be applied to arbitrary pedigree structures and arbitrary patterns of missing marker allele information.

Similar proposals have appeared previously. Kaplan et al. [1997], Cleves et al. [1997] and Lazzeroni and Lange [1998] suggest approaches to computing *p* values that involve conditioning on traits and parental marker alleles. Rabinowitz [1997] applies the conditioning approach to quantitative traits in the context of nuclear families. All of these, however, have been limited to settings where parental marker data are available. Spielman and Ewens' [1998], Knapp's [1999], and Boehnke and Langefeld's [1998] approaches using discordant siblings without parental data may be thought of as special cases of the approach advocated here, although their conditioning strategies are not uniformly equivalent to conditioning on the minimal sufficient statistic.

The key issue in using the approach is the computation of the conditional distribution of the data given the minimal sufficient statistics for the null hypothesis. In nuclear families with a single offspring and when parental marker data are available, the traits and parental marker alleles constitute the minimal sufficient statistic, and the conditional distribution of the child's marker alleles corresponds to the usual Mendelian transmission probabilities of the parental marker alleles. This corresponds to an absence of assumptions about the joint distribution of parental markers and traits in pedigree members; any joint distribution could be realized through some combination of population admixture, genetic model and ascertainment scheme. However, under the null hypotheses, regardless of patterns of population admixture, of the genetic model, or the ascertainment scheme, conditionally given the parental markers and the traits, each parent is equally likely to transmit either of her/his markers, and parental transmissions are independent. In general pedigrees, and especially in settings where not all pedigree members' markers are typed, the conditional distribution given the minimal sufficient statistic is not always immediately obvious. The traits are fixed as in the case of nuclear families with a single offspring, but the conditional distribution of the typed marker alleles is in some cases quite different from the usual Mendelian transmission probabilities.

The remainder of this paper focuses on the minimal sufficient statistic and the conditional distribution of the observed data given the minimal sufficient statistic. The

main result is an algorithm for computing the conditional distributions of the typed marker alleles, given the minimal sufficient statistic, that may be applied with arbitrary pedigree structures and arbitrary patterns of missing marker allele information. The descriptions of the conditional distributions that result when the algorithm is applied to nuclear families are also presented. The derivation of the conditional distribution relies on a fairly technical characterization of the minimal sufficient statistic. The proof of the validity of the characterization is relegated to an appendix.

In practice, the algorithm is applied separately to each distinct pedigree in a data set. Different strategies may be taken when the results of the algorithm are used to compute *p* values. The descriptions of the conditional distributions may be used to repeatedly generate, independently for each pedigree, pseudo-random marker alleles for the typed markers in the pedigrees. The test statistic may be computed for each of the resulting data sets and the resulting Monte-Carlo distribution of the test statistic may be used as the reference distribution for computing *p* values. This approach is suggested in Kaplan et al. [1997], Rabinowitz [1997], Cleves et al. [1997], Lazzeroni and Lange [1998] and Spielman and Ewens [1998], among others. Another approach to using the results of the algorithm to compute *p* values is applicable in data sets with many pedigrees and with test statistics that are the sum of pedigree-specific contributions; the expectations and variances of the conditional distributions of the pedigree-specific contributions may be derived from the conditional distribution of the observed data and then used to normalize the observed test statistic. The approximate reference distribution for the normalized statistic is then the standard Gaussian distribution. This approach is used in Spielman and Ewens [1998] and underlies an approach taken in Rabinowitz [1997], for example. A third approach to computing *p* values may be computationally feasible in data sets with only a few pedigrees: the pedigree-specific conditional distributions may be used to compute the exact conditional distribution of the test statistic through complete enumeration.

A distinction should be made between settings where investigators are using association methods to search for evidence of linkage and settings where linkage has been established in a region and association methods are being used for more precise mapping or to examine a candidate locus. The former setting might be, for example, a genome scan. The latter might be where a classical linkage analysis has identified a broad region that is then typed at a sequence of finely spaced markers, or where candidate

mutations at presumed functional loci are examined. This distinction is discussed in Martin et al. [1997] and Ewens and Spielman [1995]. The null hypothesis, and therefore the minimal sufficient statistic under the null hypothesis, are different in the two settings. In the first setting, the null hypothesis is that the marker is not linked to any trait locus. In the second, the null hypothesis is that there is independence between the alleles of the marker and the alleles of any trait locus that is linked to the marker. The two settings are treated separately in the following two sections. The fourth section details the application of the approach to a complex pedigree.

Throughout, it is assumed that traits are available for some members of pedigrees, and that in (not necessarily the same) members of the pedigrees, markers have been typed. It is assumed that a test statistic that is sensitive to association between traits and marker alleles is defined for the observed traits and typed markers. The test statistic might be as simple as restricting attention to individuals expressing a dichotomous trait and counting the number of chromosomes that carry a particular marker allele, or as complicated as a score statistic from a joint likelihood for quantitative or survival traits in all pedigree members that incorporates environmental covariates and genotype information from other loci known to influence the trait. No assumptions on the ascertainment process are made other than the usual assumption that the decision to include an individual in the study sample is made without reference to the individual's marker alleles. In each of the following two sections, the null hypothesis is first discussed and the minimal sufficient statistic under the null hypothesis for the special case where all pedigree founders' markers are typed is presented. Then, the characterization in the Appendix is used to generalize from the special case to an algorithm for the general case where not all pedigree founders' markers are typed. The results of the algorithm for nuclear families are then tabulated. A subsequent section illustrates the application of the algorithm to a nonnuclear pedigree.

A natural question that arises is whether conditioning on the minimal sufficient statistic encompasses all possible approaches to adjusting. That is, whether or not every adjustment approach either can be uniformly improved upon or can be derived through the conditioning approach. If the minimal sufficient statistic were complete, then the approach would encompass all exact tests. See, for example, Cox and Hinkley [1974]. However, with certain kinds of missing marker information, with the second type of null hypothesis, the minimal sufficient statistic is not complete. This suggests that there can be information

that is not extracted through conditioning on the minimal sufficient statistic. An approach to recovering such information is discussed in 'Recovering Information'.

Using Association Methods to Test for Linkage

In this section the setting where association methods are used to test for linkage is discussed. It is assumed that correct type I error probabilities are desired regardless of the underlying genetic model, regardless of patterns of population admixture and regardless of the sampling plan. Therefore, since any distribution of pedigree structures, traits and pedigree founder marker alleles may be obtained as the consequence of some genetic model, pattern of population admixture and sampling plan, it follows that the joint distribution of the marker alleles in pedigree founders and the observed traits in all pedigree members is unrestricted under the null hypothesis. (The term 'founder' as used here may be defined in terms of having relatives in the pedigree, but only relatives that are also descendents. Founders are individuals who satisfy this property, and also who have no descendents that satisfy the property. With this definition, marry-ins are also founders.)

When testing for linkage, the null hypothesis is that the markers are not linked to any locus that influences that trait. It follows that, conditionally given the marker alleles in founders and the observed traits in all pedigree members, under the null hypothesis, transmission of marker alleles from founders to offspring is independent of the traits in pedigree members. Therefore, if all founder marker alleles are typed, the minimal sufficient statistic under the null hypothesis would be the founder marker alleles and the observed traits in all pedigree members. The conditional distribution of the alleles at the typed markers, given the traits in the pedigree members and the founders' marker alleles, does not depend on the traits and corresponds to the usual transmission probabilities given by Mendel's laws.

When not all founders' markers are typed, the characterization in the appendix may be used to derive the minimal sufficient statistic. The characterization implies that two different outcomes (outcomes consist of all observed traits and typed marker alleles) for a pedigree correspond to the same value of the minimal sufficient statistic if and only if the two outcomes satisfy the following three conditions: first, that the traits in the two outcomes are the same; second, that any set of founder genotypes that is compatible with one of the outcomes is also compatible

Table 1. Conditional distributions when testing for linkage with one homozygous *AA* parent's marker alleles available

Children's marker alleles	Conditional distribution
1 { <i>AA</i> } or { <i>AB</i> }	observed data have conditional probability 1
2 { <i>AA</i> , <i>AB</i> }	randomly assign <i>AA</i> or <i>AB</i> with probability 1/2, 1/2, independently to each sib, discarding outcomes without at least one assignment of <i>AA</i> and one assignment of <i>AB</i>
3 { <i>AB</i> , <i>AC</i> }	randomly assign <i>AB</i> or <i>AC</i> with probability 1/2, 1/2, independently to each sib, discarding outcomes without at least one assignment of <i>AB</i> and one assignment of <i>AC</i>

with the other outcome; and third, the ratio of the conditional probabilities of the outcomes, given the founders' marker alleles, is the same for all of the compatible patterns of founder marker alleles. Note that the minimal sufficient statistic is not represented as a particular function of the data, but rather as a partition of the sample space.

An algorithm for computing the conditional distribution in a pedigree therefore takes the following form.

(1) Find all the patterns of founder marker alleles that are compatible with the alleles in the typed markers.

(2) For each of the patterns of compatible founder marker alleles obtained in the first step, find the set of compatible patterns of alleles in the typed markers in the pedigree. Find the subset of these compatible patterns that have exactly the same compatible patterns of founder marker alleles as the observed alleles in the typed markers.

(3) For every pattern of compatible founder marker alleles found in the first step and for every pattern of alleles in the typed markers in the subset found in the second step, compute the ratio of the conditional probability of the alleles in the typed markers given the pattern of founder marker alleles to the conditional probability of the observed alleles in the typed markers given the pattern of founder marker alleles.

(4) For some patterns of alleles in the typed markers in the subset found in the second step, the ratios found in the third step will be the same for all of the compatible patterns of founder marker alleles found in the first step. This subset is the set of outcomes with positive conditional probability.

(5) The conditional distribution is found by arbitrarily choosing any of the compatible patterns of founder mark-

Table 2. Conditional distributions when testing for linkage with one heterozygous *AB* parent's marker alleles available

Children's marker alleles	Conditional distribution
1 { <i>AA</i> } or { <i>AB</i> }	observed data have conditional probability 1
2 { <i>AA</i> , <i>AB</i> }	random assignment of <i>AA</i> and <i>AB</i> that keeps invariant the number of each
3 { <i>AA</i> , <i>BB</i> } or { <i>AA</i> , <i>AB</i> , <i>BB</i> }	randomly assign <i>AA</i> , <i>AB</i> and <i>BB</i> with probabilities 1/4, 1/2, 1/4, independently to each sib, discarding outcome without at least one assignment of <i>AA</i> and one assignment of <i>BB</i>
4 { <i>AC</i> } or { <i>AC</i> , <i>BC</i> }	randomly assign <i>AC</i> and <i>BC</i> with probabilities, 1/2, 1/2, independently to each sib
5 { <i>AB</i> , <i>AC</i> } or { <i>AB</i> , <i>AC</i> , <i>BC</i> }	randomly assign <i>AB</i> , <i>AC</i> and <i>BC</i> with probabilities 1/3, 1/3, 1/3 independently to each sib, discarding outcome without <i>AB</i> assigned at least once and without at least one of <i>AC</i> and <i>BC</i> assigned at least once
6 { <i>AA</i> , <i>AC</i> }, { <i>AA</i> , <i>BC</i> }, { <i>AA</i> , <i>AB</i> , <i>AC</i> }, { <i>AA</i> , <i>AB</i> , <i>BC</i> } or { <i>AA</i> , <i>AC</i> , <i>BC</i> }	randomly assign <i>AA</i> , <i>AC</i> , <i>AB</i> and <i>BC</i> with probabilities 1/4, 1/4, 1/4, 1/4, discarding outcomes without <i>AA</i> assigned at least once and without at least one of <i>AC</i> and <i>BC</i> assigned at least once
7 { <i>AC</i> , <i>BD</i> }, { <i>AC</i> , <i>AD</i> }, { <i>AC</i> , <i>BC</i> , <i>BD</i> } or { <i>AC</i> , <i>BC</i> , <i>BD</i> , <i>AD</i> }	randomly assign <i>AC</i> , <i>AD</i> , <i>BC</i> and <i>BD</i> with probabilities 1/4, 1/4, 1/4, 1/4, discarding outcomes in which either <i>C</i> or <i>D</i> are not assigned

er alleles found in the first step and computing the conditional probabilities of the patterns of alleles in the typed markers given the chosen pattern of founders' marker alleles and given that the markers lie in the set described in the fourth step.

The conditional distributions for configurations of observed data in nuclear families are presented in tables 1–3. Table 1 is for the case of marker alleles observed in only a single homozygous parent. Table 2 is for observed marker alleles in only a single heterozygous parent. Table 3 is for the case where marker alleles are not observed in either parent. The tables have two columns.

The first column corresponds to the observed configuration of marker alleles in the children. Throughout, *A*, *B*, *C* and *D* are used as generic marker alleles. The configurations are listed in the form of sets. The notation {*AB*, *AC*} in the third entry in table 1, for example, corresponds to a sibship of arbitrary size with at least one child carrying *AB*

Table 3. Conditional distributions when testing for linkage with no parent's marker alleles available

Children's marker alleles	Conditional distribution
1 { <i>AA</i> } or { <i>AB</i> }	observed data have conditional probability 1
2 { <i>AA</i> , <i>AB</i> }	random assignment of <i>AA</i> and <i>AB</i> that keeps invariant the number of each
3 { <i>AA</i> , <i>BB</i> } or { <i>AA</i> , <i>AB</i> , <i>BB</i> }	randomly assign <i>AA</i> , <i>BB</i> and <i>AB</i> with probabilities 1/4, 1/2, 1/4, independently to each sib, discarding outcome without at least one assignment of <i>AA</i> and one assignment of <i>BB</i>
4 { <i>AB</i> , <i>AC</i> , <i>BC</i> }	randomly assign <i>AB</i> , <i>AC</i> and <i>BC</i> with probabilities 1/3, 1/3, 1/3 independently to each sib, discarding outcome without <i>AB</i> , <i>AC</i> and <i>BC</i> each assigned at least once
5 { <i>AB</i> , <i>AC</i> }	randomly assign <i>AB</i> and <i>AC</i> with probabilities 1/2, 1/2, independently to each sib, discarding outcome without <i>AB</i> and <i>AC</i> assigned at least once
6 { <i>AA</i> , <i>BC</i> }, { <i>AA</i> , <i>AB</i> , <i>AC</i> }, { <i>AA</i> , <i>AC</i> , <i>BC</i> } or { <i>AA</i> , <i>AB</i> , <i>AC</i> , <i>BC</i> }	randomly assign <i>AA</i> , <i>AB</i> , <i>AC</i> and <i>BC</i> with probabilities 1/4, 1/4, 1/4, 1/4, discarding outcomes without <i>AA</i> assigned at least once and outcomes without both <i>B</i> and <i>C</i> represented
7 { <i>AC</i> , <i>BD</i> }	randomly assign <i>AC</i> and <i>BD</i> with equal probabilities, independently to each sib, discarding outcome without at least one assignment of <i>AC</i> and one assignment of <i>BD</i>
8 { <i>AC</i> , <i>BC</i> , <i>BD</i> } or { <i>AC</i> , <i>BC</i> , <i>AD</i> , <i>BD</i> }	randomly assign <i>AC</i> , <i>BC</i> , <i>AD</i> , <i>BD</i> with equal probabilities, discarding outcomes that do not contain at least three of the four

and one child carrying *AC* and no other genotypes represented in the sibship. More precisely, because *A*, *B*, *C* and *D* refer to generic markers, the notation {*AB*, *AC*} corresponds to a sibship of arbitrary size with the same marker allele (referred to by *A*) carried by all children and exactly two other marker alleles (referred to by *B* and *C*) represented in the sibship.

The second column is a description of the conditional distribution in the form of an algorithm for simulating the distribution. The description

Randomly assign *AB* or *AC* with probability 1/2, 1/2, independently to each sib, discarding outcomes without at least one assignment of *AB* and one assignment of *AC*.

in the third entry in table 1, for example, means that the conditional distribution may be simulated by randomly

assigning AB or AC with equal probability, independently in each child, until a configuration of assignments occurs that has at least one assignment of AB and one assignment of AC .

Testing for Association in the Presence of Linkage

In this section, the setting of testing a candidate mutation or using association for fine mapping in a region where linkage has been established is treated. As in the setting of the previous section, it is assumed that correct type I error probabilities are desired for any genetic model, for any pattern of population admixture and for any sampling plan. The null hypothesis, therefore, again leaves unspecified the joint distribution of traits and the marker alleles in the pedigree founders.

Unlike the setting of the previous section, however, the null hypothesis does not specify that the marker is not linked to any locus that influences the trait. Conditionally given marker alleles in founders, therefore, identity-by-descent relationships (identity-by-descent relationships are the patterns of allele sharing, not simply the count of the alleles shared by pedigree members) are not necessarily independent of the pattern of traits in a pedigree. This is because identity-by-descent relationships for the marker alleles may be associated with transmission of linked loci that influence the trait, which is in turn associated with the traits in the pedigree. It follows that the joint distribution of traits, founder marker alleles and identity-by-descent relationships is unspecified under the null hypothesis.

The null hypothesis does, however, specify that, in the population from which the pedigrees are chosen, there is no association between the marker alleles and any linked locus that influences the trait. Therefore, in settings where founder genotypes are observed and where identity-by-descent relationships are uniformly available as well, the minimal sufficient statistic would be the traits, founder marker alleles and identity-by-descent relationships. The conditional distribution of transmissions given the minimal sufficient statistic would correspond to transmission of marker alleles independently of the traits and in accordance with Mendel's laws restricted to outcomes that preserve the identity-by-descent relationships. In nuclear families where both parents' marker alleles are observed and where identity by descent relationships are given (for example, through typing flanking markers) the conditional distribution is generated by randomly choosing neither, both or one or the other of the parents and switching in all

offspring, the transmitted marker allele from the chosen parents. See, for example, Martin et al. [1997].

In order to generalize to the setting where founder marker alleles may not be typed or where identity-by-descent relationships are not available, the characterization in the appendix may again be used. The application of the characterization involves, in this case, incorporation of the identity-by-descent relationships. Specifically, in the second of the three conditions given in the previous section, the compatible patterns of founders' marker alleles must be augmented to include compatible identity-by-descent relationships as well. In the third condition, the conditional probabilities of alleles in the typed markers, given the compatible parental markers should be replaced by the conditional probabilities given the compatible parental markers and the compatible identity-by-descent relationships. It follows that the description of the

Table 4. Conditional distributions when testing for association in the presence of linkage with both parents' marker alleles available

Parent's marker alleles	Conditional distribution
1 $AA\ AA$	observed data have conditional probability 1
2 $AA\ AB$	$AA \rightarrow AB, AB \rightarrow AA$, w.p. 1/2, and no change, w.p. 1/2
3 $AA\ BB$	observed data have conditional probability 1
4 $AA\ BC$	$AB \rightarrow AC, AC \rightarrow AB$, w.p. 1/2, and no change, w.p. 1/2
5 $AB\ AB$	$AA \rightarrow BB, BB \rightarrow AA$, w.p. 1/2, and no change, w.p. 1/2, except in the case of exactly one heterozygous and one homozygous offspring; in that case, for
5'	$\{AA, AB\}, AA \rightarrow AB, AB \rightarrow AA$ w.p. 1/4, $AA \rightarrow AB, AB \rightarrow BB$ w.p. 1/4, $AA \rightarrow BB, AB \rightarrow AB$ w.p. 1/4, and no change w.p. 1/4, and, for
5''	$\{BB, AB\}, BB \rightarrow AB, AB \rightarrow BB$ w.p. 1/4, $BB \rightarrow AB, AB \rightarrow AA$ w.p. 1/4, $BB \rightarrow AA, AB \rightarrow AB$ w.p. 1/4, and no change w.p. 1/4
6 $AB\ AC$	$AA \rightarrow BA, AC \rightarrow BC, AB \rightarrow AA, BC \rightarrow AC$, w.p. 1/4, $AA \rightarrow AC, AC \rightarrow AA, AB \rightarrow BC, BC \rightarrow AB$, w.p. 1/4, $AA \rightarrow BC, AC \rightarrow AB, AB \rightarrow AC, BC \rightarrow AA$, w.p. 1/4, and no change, w.p. 1/4
7 $AC\ BD$	$AB \rightarrow BC, AD \rightarrow CD, BC \rightarrow AB, CD \rightarrow AD$, w.p. 1/4, $AB \rightarrow AD, AD \rightarrow AB, BC \rightarrow CD, CD \rightarrow BC$, w.p. 1/4, $AB \rightarrow CD, AD \rightarrow CB, BC \rightarrow AD, CD \rightarrow AB$, w.p. 1/4, and no change, w.p. 1/4

algorithm given in the previous section should be modified by appending the phrase ‘and identity-by-descent relationships’ to each appearance of ‘patterns of compatible founder marker alleles’.

The conditional distributions for nuclear families for testing for association in the presence of linkage are listed in tables 4–7. Table 4 is for the case where marker alleles are observed in both parents. Tables 5–7 are for a single homozygous parent, a single heterozygous parent and for no observed parental markers, respectively. Table 4 differs from the other tables in that the first column corresponds to marker alleles in the parents rather than in the

Table 5. Conditional distributions when testing for association in the presence of linkage with one homozygous *AA* parent’s marker alleles available

Children’s marker alleles	Conditional distribution
1 { <i>AA</i> } or { <i>AB</i> }	observed data have conditional probability 1
2 { <i>AA</i> , <i>AB</i> }	<i>AA</i> → <i>AB</i> , <i>AB</i> → <i>AA</i> , w.p. 1/2, no change w.p. 1/2
3 { <i>AB</i> , <i>AC</i> }	<i>AB</i> → <i>AC</i> , <i>AC</i> → <i>AB</i> , w.p. 1/2, no change w.p. 1/2

Table 6. Conditional distributions when testing for association in the presence of linkage with one heterozygous *AB* parent’s marker alleles available

Children’s marker alleles	Conditional distribution
1 { <i>AA</i> }, { <i>AB</i> } or { <i>AB</i> , <i>AC</i> , <i>BC</i> }	observed data have conditional probability 1
2 { <i>AA</i> , <i>AB</i> }	observed data have conditional probability 1, except if only two offspring; then, <i>AA</i> → <i>AB</i> , <i>AB</i> → <i>AA</i> , w.p. 1/2, and no change w.p. 1/2
3 { <i>AA</i> , <i>BB</i> } or { <i>AA</i> , <i>AB</i> , <i>BB</i> }	<i>AA</i> → <i>BB</i> , <i>BB</i> → <i>AA</i> , w.p. 1/2, and no change w.p. 1/2
4 { <i>AC</i> }, { <i>AC</i> , <i>BC</i> }	<i>AC</i> → <i>BC</i> , <i>BC</i> → <i>AC</i> , w.p. 1/2, and no change w.p. 1/2
5 { <i>AB</i> , <i>AC</i> }	<i>AB</i> → <i>AC</i> , <i>AC</i> → <i>AB</i> , w.p. 1/2, and no change w.p. 1/2
6 { <i>AA</i> , <i>AC</i> }	<i>AA</i> → <i>AC</i> , <i>AC</i> → <i>AA</i> , w.p. 1/2, and no change w.p. 1/2
7 { <i>AA</i> , <i>BC</i> }	<i>AA</i> → <i>BC</i> , <i>BC</i> → <i>AA</i> , w.p. 1/2, and no change w.p. 1/2
8 { <i>AA</i> , <i>AB</i> , <i>AC</i> }	<i>AA</i> → <i>AB</i> , <i>AB</i> → <i>AA</i> , <i>AC</i> → <i>BC</i> , w.p. 1/3, <i>AA</i> → <i>AC</i> , <i>AB</i> → <i>BC</i> , <i>AC</i> → <i>AA</i> , w.p. 1/3, and no change w.p. 1/3
9 { <i>AA</i> , <i>AB</i> , <i>BC</i> }	<i>AA</i> → <i>AB</i> , <i>AB</i> → <i>AA</i> , <i>BC</i> → <i>AC</i> , w.p. 1/3, <i>AA</i> → <i>BC</i> , <i>AB</i> → <i>AC</i> , <i>BC</i> → <i>AA</i> , w.p. 1/3, and no change w.p. 1/3
10 { <i>AA</i> , <i>AC</i> , <i>BC</i> }	<i>AA</i> → <i>BC</i> , <i>AC</i> → <i>AB</i> , <i>BC</i> → <i>AA</i> , w.p. 1/3, <i>AA</i> → <i>BC</i> , <i>AC</i> → <i>AB</i> , <i>BC</i> → <i>AC</i> , w.p. 1/3, and no change w.p. 1/3
11 { <i>AA</i> , <i>AC</i> , <i>AB</i> , <i>BC</i> }	<i>AA</i> → <i>AB</i> , <i>AC</i> → <i>BC</i> , <i>AB</i> → <i>AA</i> , <i>BC</i> → <i>AC</i> , w.p. 1/4, <i>AA</i> → <i>AC</i> , <i>AC</i> → <i>AA</i> , <i>AB</i> → <i>BC</i> , <i>BC</i> → <i>AB</i> , w.p. 1/4, <i>AA</i> → <i>BC</i> , <i>AC</i> → <i>AB</i> , <i>AB</i> → <i>AC</i> , <i>BC</i> → <i>AA</i> , w.p. 1/4, and no change, w.p. 1/4
12 { <i>AC</i> , <i>BD</i> }, { <i>AC</i> , <i>AD</i> }, { <i>AC</i> , <i>BC</i> , <i>BD</i> } or { <i>AC</i> , <i>BC</i> , <i>BD</i> , <i>AD</i> }	<i>AC</i> → <i>BD</i> , <i>BD</i> → <i>AD</i> , <i>BC</i> → <i>AC</i> , <i>AD</i> → <i>BD</i> , w.p. 1/4, <i>AC</i> → <i>AD</i> , <i>BD</i> → <i>AC</i> , <i>BC</i> → <i>BD</i> , <i>AD</i> → <i>AC</i> , w.p. 1/4, <i>AC</i> → <i>BD</i> , <i>BD</i> → <i>AC</i> , <i>BC</i> → <i>AD</i> , <i>AD</i> → <i>BC</i> , w.p. 1/4, and no change, w.p. 1/4

Table 7. Conditional distributions when testing for association in the presence of linkage with no parent's marker alleles available

Children's marker alleles	Conditional distribution
1 {AA}, {AB} or {AB, AC, BC}	observed data have conditional probability 1
2 {AA, AB} 2'	observed data have conditional probability 1, except if only two offspring; then, $AA \rightarrow AB, AB \rightarrow AA$, w.p. 1/2 and no change w.p. 1/2
3 {AA, BB} or {AA, AB, BB}	$AA \rightarrow BB, BB \rightarrow AA$, w.p. 1/2 and no change w.p. 1/2
4 {AB, AC}	$AB \rightarrow AC, AC \rightarrow AB$, w.p. 1/2 and no change w.p. 1/2
5 {AA, BC}	$AA \rightarrow BC, BC \rightarrow AA$, w.p. 1/2 and no change w.p. 1/2
6 {AC, BD}	$AC \rightarrow BD, BD \rightarrow AC$, w.p. 1/2 and no change w.p. 1/2
7 {AA, AB, AC}	$AA \rightarrow AB, AB \rightarrow AA, AC \rightarrow BC$, w.p. 1/3, $AA \rightarrow AC, AB \rightarrow BC, AC \rightarrow AA$, w.p. 1/3 and no change w.p. 1/3
8 {AA, AC, BC}	$AA \rightarrow AC, AC \rightarrow AA, BC \rightarrow AB$, w.p. 1/3, $AA \rightarrow BC, AC \rightarrow AB, BC \rightarrow AA$, w.p. 1/3 and no change w.p. 1/3
9 {AA, AB, AC, BC}	$AA \rightarrow AB, AB \rightarrow AA, AC \rightarrow BC, BC \rightarrow AC$, w.p. 1/4, $AA \rightarrow AC, AB \rightarrow BC, AC \rightarrow AA, BC \rightarrow AB$, w.p. 1/4, $AA \rightarrow BC, AB \rightarrow AC, AC \rightarrow AB, BC \rightarrow AA$, w.p. 1/4 and no change w.p. 1/4
10 {AC, BC, BD} or {AC, BC, BD, AD}	$AC \rightarrow BC, BC \rightarrow AC, AD \rightarrow BD, BD \rightarrow AD$, w.p. 1/4, $AC \rightarrow AD, BC \rightarrow BD, AD \rightarrow AC, BD \rightarrow BC$, w.p. 1/4, $AC \rightarrow BD, BC \rightarrow AD, AD \rightarrow BC, BD \rightarrow AC$, w.p. 1/4 and no change w.p. 1/4

children. In the descriptions of algorithms for simulating the conditional distributions in the second columns of the tables, the description

$AA \rightarrow AB, AB \rightarrow AA$, w.p. 1/2.

No change, w.p. 1/2

from the second entry in table 4, for example, means that the conditional distribution is generated by leaving the data unchanged with probability 1/2 and interchanging the genotypes in the children with probability 1/2.

Application to a Nonnuclear Pedigree

In this section, the algorithm is illustrated through an application to the pedigree depicted in figure 1. The lower-case roman numerals in the figure index individuals in

the pedigree. Individuals *i*, *ii*, *iii* and *vi* are the founders. Missing marker alleles are indicated with question marks. All founders' marker alleles are missing. First, the case of using association methods to search for evidence of linkage is presented. Then, the case of using association methods for more precise mapping or to examine a candidate locus in the presence of linkage is presented.

The first step of the algorithm in the first case results in three possible patterns of marker alleles in the founders compatible with the observed data. These three patterns are listed in table 8. The entries in the table are marker genotypes. The four columns of the table correspond to the four founders. Individuals *iii* and *vi* are obligate heterozygous *AB*. More than one pattern of compatible marker alleles exist for individuals *i* and *ii*.

The patterns of alleles in the typed markers that are compatible with exactly the same set of patterns of marker

Table 8. Patterns of founder marker in the example

<i>i</i>	<i>ii</i>	<i>iii</i>	<i>vi</i>
<i>AA</i>	<i>AB</i>	<i>AB</i>	<i>AB</i>
<i>AB</i>	<i>AA</i>	<i>AB</i>	<i>AB</i>
<i>AB</i>	<i>AB</i>	<i>AB</i>	<i>AB</i>

Table 9. Patterns of alleles in the typed markers compatible with exactly the same patterns of compatible founder marker alleles as the observed marker alleles

<i>iv</i>	<i>v</i>	<i>vii</i>	<i>viii</i>	<i>ix</i>	<i>x</i>
<i>AB</i>	<i>AA</i>	<i>BB</i>	<i>AA</i>	<i>AA</i>	<i>AB</i>
<i>AB</i>	<i>AA</i>	<i>AA</i>	<i>BB</i>	<i>AA</i>	<i>AB</i>
<i>AB</i>	<i>AA</i>	<i>BB</i>	<i>AA</i>	<i>AB</i>	<i>AA</i>
<i>AB</i>	<i>AA</i>	<i>AA</i>	<i>BB</i>	<i>AB</i>	<i>AA</i>
<i>AA</i>	<i>AB</i>	<i>AA</i>	<i>AB</i>	<i>BB</i>	<i>AA</i>
<i>AA</i>	<i>AB</i>	<i>AA</i>	<i>AB</i>	<i>AA</i>	<i>BB</i>
<i>AA</i>	<i>AB</i>	<i>AB</i>	<i>AA</i>	<i>BB</i>	<i>AA</i>
<i>AA</i>	<i>AB</i>	<i>AB</i>	<i>AA</i>	<i>AA</i>	<i>BB</i>

alleles in the founders as the observed alleles in the typed markers are listed in table 9. The entries in the table are genotypes and the columns correspond to the six individuals whose markers are typed. The second step of the algorithm results in eight patterns.

For the first and second of the patterns of founder marker alleles found in the first step of the algorithm, the conditional probability of all of the marker alleles in typed individuals found in the second stage are $1/2^8$, while for the third, the conditional probabilities are all $1/2^9$. The ratios are all equal to 1 for all three patterns of founder marker alleles. The subset from the fourth step of the algorithm is therefore the whole set of eight patterns of marker alleles in the typed individuals. Furthermore, the conditional distribution gives equal probability to all eight patterns.

Now, consider the second case, testing for association in the presence of linkage. Table 10 lists the patterns of founder marker alleles and identity-by-descent relationships compatible with the observed data. The rows in the table correspond to different possible patterns of founder marker alleles and identity-by-descent relationships. The first four columns are the marker genotypes in the founders. The next eight columns list the indices of individuals

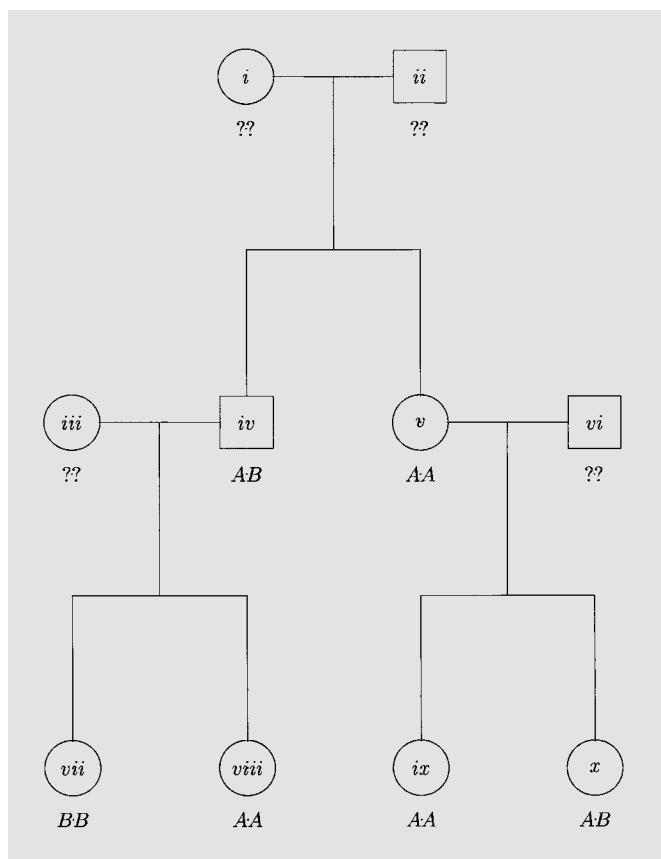


Fig. 1. Pedigree used in the example to illustrate the approach.

sharing founder marker alleles. Each founder potentially contributes two alleles. Individual *ii*'s two alleles, for example, are distinguished in the columns by the notation *ii* and *ii'*.

Only the first and third rows of table 9 correspond to outcomes with exactly the same set of compatible founder marker alleles and identity-by-descent relationships as the observed data. The first row is the observed data and the third row corresponds to interchanging the alleles of individual *vi*. Paradoxically, although subjects *iii*, *iv*, *vii* and *viii* taken as a pedigree by themselves would provide unbiased information that is recovered through the random assignment of *BB* to *vii* and *AA* to *viii* or *AA* to *vii* and *BB* to *viii*, the interchange is not allowed under the conditional distribution given the minimal sufficient statistic. This is because, for example, it is possible that subject *i* carries *AA* and subject *ii* carries *AB*, and one of subject *i*'s marker alleles is shared identical by descent with subjects *viii*, *ix* and *x*. This scenario is incompatible with subject *vii* being homozygous *AA* and subject *viii* being homozy-

gous *BB*. For all of the outcomes other than first and third, the compatible identity by descent relationships are similarly not the same as those compatible with the observed data. The two allowed outcomes are equally likely under the conditional distribution.

It is instructive to examine how the calculation of the conditional distribution given the sufficient statistic may be used to compute the expectation and variance of a test statistic. Let T_{iv} , T_v , T_{vii} , T_{viii} , T_{ix} and T_x denote traits in the individuals with typed markers. The traits might be indicators of affected status or residuals from the regression of quantitative traits on covariates. Let Y_{iv} , Y_v , Y_{vii} , Y_{viii} , Y_{ix} and Y_x be variables that reflect the number of a particular allele carried by the individuals with typed markers. The variables might simply count the number of the alleles carried or be some other function of the number of alleles. Let y_0 , y_1 and y_2 denote the values of the function for zero, one or two alleles. Consider a test statistic of the form

$$T_{iv}Y_{iv} + T_vY_v + T_{vii}Y_{vii} + T_{viii}Y_{viii} + T_{ix}Y_{ix} + T_xY_x.$$

Most test statistics that have appeared in the literature take the form of adjusted versions of such test statistics.

When testing for linkage, the conditional expectation under the null hypothesis of the test statistic is

$$\begin{aligned} \mu = & \frac{1}{8} (T_{iv}y_1 + T_vy_2 + T_{vii}y_0 + T_{viii}y_2 + T_{ix}y_2 + T_xy_1 + \\ & T_{iv}y_1 + T_vy_2 + T_{vii}y_2 + T_{viii}y_0 + T_{ix}y_2 + T_xy_1 + \\ & T_{iv}y_1 + T_vy_2 + T_{vii}y_0 + T_{viii}y_2 + T_{ix}y_1 + T_xy_2 + \\ & T_{iv}y_1 + T_vy_2 + T_{vii}y_2 + T_{viii}y_0 + T_{ix}y_1 + T_xy_2 + \\ & T_{iv}y_2 + T_vy_1 + T_{vii}y_2 + T_{viii}y_1 + T_{ix}y_0 + T_xy_2 + \\ & T_{iv}y_2 + T_vy_1 + T_{vii}y_2 + T_{viii}y_1 + T_{ix}y_2 + T_xy_0 + \\ & T_{iv}y_2 + T_vy_1 + T_{vii}y_1 + T_{viii}y_2 + T_{ix}y_0 + T_xy_2 + \\ & T_{iv}y_2 + T_vy_1 + T_{vii}y_1 + T_{viii}y_2 + T_{ix}y_2 + T_xy_0). \end{aligned}$$

The conditional variance under the null hypothesis is

$$\begin{aligned} & \frac{1}{8} ((T_{iv}y_1 + T_vy_2 + T_{vii}y_0 + T_{viii}y_2 + T_{ix}y_2 + T_xy_1 - \mu)^2 + \\ & (T_{iv}y_1 + T_vy_2 + T_{vii}y_2 + T_{viii}y_0 + T_{ix}y_2 + T_xy_1 - \mu)^2 + \\ & (T_{iv}y_1 + T_vy_2 + T_{vii}y_0 + T_{viii}y_2 + T_{ix}y_1 + T_xy_2 - \mu)^2 + \\ & (T_{iv}y_1 + T_vy_2 + T_{vii}y_2 + T_{viii}y_0 + T_{ix}y_1 + T_xy_2 - \mu)^2 + \\ & (T_{iv}y_2 + T_vy_1 + T_{vii}y_2 + T_{viii}y_1 + T_{ix}y_0 + T_xy_2 - \mu)^2 + \\ & (T_{iv}y_2 + T_vy_1 + T_{vii}y_2 + T_{viii}y_1 + T_{ix}y_2 + T_xy_0 - \mu)^2 + \\ & (T_{iv}y_2 + T_vy_1 + T_{vii}y_1 + T_{viii}y_2 + T_{ix}y_0 + T_xy_2 - \mu)^2 + \\ & (T_{iv}y_2 + T_vy_1 + T_{vii}y_1 + T_{viii}y_2 + T_{ix}y_2 + T_xy_0 - \mu)^2). \end{aligned}$$

To take a concrete example, suppose that the traits are indicators of affected status, with 1 indicating affected

and 0 indicating unaffected status, and suppose that the statistic counts the number of *A* alleles, so that $y_0 = 0$, $y_1 = 1$ and $y_2 = 2$ according to how many *A* alleles are carried. Suppose that subjects *ix* and *x* are affected and all other subjects are not. Then, the contribution from the pedigree to the unadjusted statistic is 3. The conditional expectation of the contribution is

$$\begin{aligned} & \frac{1}{8} ((2 + 1) + (2 + 1) + (1 + 2) + (1 + 2) + (0 + 2) + (2 + 0) + \\ & (0 + 2) + (2 + 0)) = \frac{20}{8} = 2.5. \end{aligned}$$

The contribution to the variance of the statistic is

$$\begin{aligned} & \frac{1}{8} ((2 + 1 - 20/8)^2 + (2 + 1 - 20/8)^2 + \\ & (1 + 2 - 20/8)^2 + (1 + 2 - 20/8)^2 + (0 + 2 - 20/8)^2 + \\ & (2 + 0 - 20/8)^2 + (0 + 2 - 20/8)^2 + (2 + 0 - 20/8)^2) = 0.25. \end{aligned}$$

Note that for this statistic, for this configuration of traits, none of the three nuclear families that make up the pedigree provide, by themselves, any information. In the nuclear family with parents *i* and *ii* and the nuclear family with parents *iii* and *iv* there are no affected individuals so the test statistic is zero and has variance zero. For the family with parents *v* and *vi*, using line 2 of table 1, the two genotypes *AA* and *AB* are randomly interchanged with probability 1/2 in the two affected individuals. Conditionally, the test statistic is constant. Under the conditional distribution in the analysis of the whole pedigree, individual *v* has genotype *AA* or *AB* with probabilities 1/2. When individual *v* has genotype *AA*, one child carries *AA* and the other *AB*. However, when individual *v* has genotype *AB*, one child has genotype *AA* and the other has *BB*. In the analysis of the nuclear family by itself, individual *v*'s marker alleles are fixed at *AA*.

Recovering Information

Conditioning on the minimal sufficient statistic is the broadest conditioning strategy for which the resulting conditional distribution is invariant to the distributions in the null hypothesis. That is, it is most efficient among all conditioning strategies that allow the conditional distribution under the null hypothesis of all test statistics to be computed exactly. However, it can be that not all the available information is captured through the conditioning approach.

This would be the case in the example of the previous section when testing for association in the presence of linkage if, for example, one of individuals *vii* and *viii* were

Table 10. Identity-by-descent relationships and founder marker alleles compatible with typed markers in the example

Founder Markers				Allele sharing of founder chromosomes							
<i>i</i>	<i>ii</i>	<i>iii</i>	<i>vi</i>	<i>i</i>	<i>i'</i>	<i>ii</i>	<i>ii'</i>	<i>iii</i>	<i>iii'</i>	<i>vi</i>	<i>vi'</i>
<i>AA</i>	<i>AB</i>	<i>AB</i>	<i>AB</i>	<i>iv,v, viii, ix, x</i>		<i>v</i>	<i>iv, vii</i>	<i>viii</i>	<i>vii</i>	<i>ix</i>	<i>x</i>
<i>AA</i>	<i>AB</i>	<i>AB</i>	<i>AB</i>	<i>iv, v, viii, ix</i>		<i>v, x</i>	<i>iv, vii</i>	<i>viii</i>	<i>vii</i>	<i>ix</i>	<i>x</i>
<i>AA</i>	<i>AB</i>	<i>AB</i>	<i>AB</i>	<i>iv, v, viii, x</i>		<i>v, ix</i>	<i>iv, vii</i>	<i>viii</i>	<i>vii</i>	<i>ix</i>	<i>x</i>
<i>AA</i>	<i>AB</i>	<i>AB</i>	<i>AB</i>	<i>iv, v, viii</i>		<i>v, ix, x</i>	<i>iv, vii</i>	<i>viii</i>	<i>vii</i>	<i>ix</i>	<i>x</i>
<i>AA</i>	<i>AB</i>	<i>AB</i>	<i>AB</i>	<i>iv, viii</i>	<i>v, ix, x</i>	<i>v</i>	<i>iv, vii</i>	<i>viii</i>	<i>vii</i>	<i>ix</i>	<i>x</i>
<i>AA</i>	<i>AB</i>	<i>AB</i>	<i>AB</i>	<i>iv, viii</i>	<i>v, ix</i>	<i>v, x</i>	<i>iv, vii</i>	<i>viii</i>	<i>vii</i>	<i>ix</i>	<i>x</i>
<i>AA</i>	<i>AB</i>	<i>AB</i>	<i>AB</i>	<i>iv, viii</i>	<i>v, x</i>	<i>v, ix</i>	<i>iv, vii</i>	<i>viii</i>	<i>vii</i>	<i>ix</i>	<i>x</i>
<i>AA</i>	<i>AB</i>	<i>AB</i>	<i>AB</i>	<i>iv, viii</i>	<i>v</i>	<i>v, ix, x</i>	<i>iv, vii</i>	<i>viii</i>	<i>vii</i>	<i>ix</i>	<i>x</i>
<i>AB</i>	<i>AA</i>	<i>AB</i>	<i>AB</i>	<i>v</i>	<i>iv, vii</i>	<i>iv, v, viii, ix, x</i>		<i>viii</i>	<i>vii</i>	<i>ix</i>	<i>x</i>
<i>AB</i>	<i>AA</i>	<i>AB</i>	<i>AB</i>	<i>v, x</i>	<i>iv, vii</i>	<i>iv, v, viii, ix</i>		<i>viii</i>	<i>vii</i>	<i>ix</i>	<i>x</i>
<i>AB</i>	<i>AA</i>	<i>AB</i>	<i>AB</i>	<i>v, ix</i>	<i>iv, vii</i>	<i>iv, v, viii, x</i>		<i>viii</i>	<i>vii</i>	<i>ix</i>	<i>x</i>
<i>AB</i>	<i>AA</i>	<i>AB</i>	<i>AB</i>	<i>v, ix, x</i>	<i>iv, vii</i>	<i>iv, v, viii</i>		<i>viii</i>	<i>vii</i>	<i>ix</i>	<i>x</i>
<i>AB</i>	<i>AA</i>	<i>AB</i>	<i>AB</i>	<i>v</i>	<i>iv, vii</i>	<i>iv, viii</i>	<i>v, ix, x</i>	<i>viii</i>	<i>vii</i>	<i>ix</i>	<i>x</i>
<i>AB</i>	<i>AA</i>	<i>AB</i>	<i>AB</i>	<i>v, x</i>	<i>iv, vii</i>	<i>iv, viii</i>	<i>v, ix</i>	<i>viii</i>	<i>vii</i>	<i>ix</i>	<i>x</i>
<i>AB</i>	<i>AA</i>	<i>AB</i>	<i>AB</i>	<i>v, ix</i>	<i>iv, vii</i>	<i>iv, viii</i>	<i>v, x</i>	<i>viii</i>	<i>vii</i>	<i>ix</i>	<i>x</i>
<i>AB</i>	<i>AA</i>	<i>AB</i>	<i>AB</i>	<i>v, ix, x</i>	<i>iv, vii</i>	<i>iv, viii</i>	<i>v</i>	<i>viii</i>	<i>vii</i>	<i>ix</i>	<i>x</i>
<i>AB</i>	<i>AB</i>	<i>AB</i>	<i>AB</i>	<i>iv, v, viii, ix, x</i>		<i>v</i>	<i>iv, vii</i>	<i>viii</i>	<i>vii</i>	<i>ix</i>	<i>x</i>
<i>AB</i>	<i>AB</i>	<i>AB</i>	<i>AB</i>	<i>iv, v, viii, ix</i>		<i>v, x</i>	<i>iv, vii</i>	<i>vii</i>	<i>vii</i>	<i>ix</i>	<i>x</i>
<i>AB</i>	<i>AB</i>	<i>AB</i>	<i>AB</i>	<i>iv, v, viii, x</i>		<i>v, ix</i>	<i>iv, vii</i>	<i>viii</i>	<i>vii</i>	<i>ix</i>	<i>x</i>
<i>AB</i>	<i>AB</i>	<i>AB</i>	<i>AB</i>	<i>iv, v, viii</i>		<i>v, ix, x</i>	<i>iv, vii</i>	<i>viii</i>	<i>vii</i>	<i>ix</i>	<i>x</i>
<i>AB</i>	<i>AB</i>	<i>AB</i>	<i>AB</i>	<i>v, ix, x</i>	<i>iv, vii</i>	<i>iv, v, viii</i>		<i>viii</i>	<i>vii</i>	<i>ix</i>	<i>x</i>
<i>AB</i>	<i>AB</i>	<i>AB</i>	<i>AB</i>	<i>v, ix</i>	<i>iv, vii</i>	<i>iv, v, viii, x</i>		<i>viii</i>	<i>vii</i>	<i>ix</i>	<i>x</i>
<i>AB</i>	<i>AB</i>	<i>AB</i>	<i>AB</i>	<i>v, x</i>	<i>iv, vii</i>	<i>iv, v, viii, ix</i>		<i>viii</i>	<i>vii</i>	<i>ix</i>	<i>x</i>
<i>AB</i>	<i>AB</i>	<i>AB</i>	<i>AB</i>	<i>v</i>	<i>iv, vii</i>	<i>iv, v, viii, ix, x</i>		<i>viii</i>	<i>vii</i>	<i>ix</i>	<i>x</i>

affected and the other were not. Then, permutations of marker alleles in *vii* and *viii* would be part of the conditional distribution if the nuclear family consisting of individuals *iii*, *iv*, *vii* and *viii* were considered by itself. These permutations are not part of the conditional distribution when the whole pedigree is considered, however. Intuitively, this is because, although the conditional distribution of marker alleles in separate portions of the pedigree can be determined, the joint conditional distribution of markers in the whole pedigree cannot.

The full strength of conditioning on sufficient statistics may not be necessary, and by giving up the ability to compute exact conditional distributions for all test statistics, additional information may be recovered. For test statistics that are made up of sums of terms in which each term involves only a portion of a pedigree, each term may be normalized by subtracting the conditional expectation given the minimal sufficient statistic for only the relevant portion of the pedigree. Robust approaches akin to those used with generalized estimating equations may then be applied to estimate the variance of the normalized sums.

These variance estimates would simply be the sum, over pedigrees, of the square of the sum within pedigrees of the various normalized terms.

Current practice with family based association tests with complex pedigrees appears to break up the pedigrees into nuclear families and apply simple test statistics to each of the families. This strategy may allow the practitioner to recover information from each nuclear pedigree that might not be available through conditioning on the minimal sufficient statistic. On the other hand, this strategy does not best extract information that might only be available from considering the traits in all individuals in the pedigree. Moreover, as seen in the example in the previous section, there may be information in the marker alleles in the pedigree as a whole that is not available from the nuclear families by themselves. What test statistic might best extract information from the whole pedigrees and the settings in which the strategy of breaking complex pedigrees into nuclear families might be preferable is an open issue.

Discussion

This paper presents an approach to testing for association while avoiding bias due to admixture. When complete founder marker information is available, in the setting of testing for linkage, the adjustment involves conditional distributions that correspond to Mendelian transmissions. When testing for association in the presence of linkage, when complete founder marker information and identity-by-descent relationships are available, the approach corresponds to randomly interchanging founder marker alleles.

The approach presented here differs from previous approaches in two fundamental ways. First, the foundation for the approach is not based on treating non-transmitted alleles as controls, nor is it based on the application of permutation statistics such as McNemar's or Mantel-Haenszel statistics. Rather, the foundation for the approach is the classical and systematic statistical strategy of conditioning on minimal sufficient statistics under the null hypothesis. Second, the question of what test statistic to use is decoupled from the problem of adjusting for admixture.

By taking the foundation to be conditioning on sufficient statistics, family-based association tests are immediately generalized to arbitrary pedigree structures and arbitrary patterns of missing marker information. By focusing on minimal sufficient statistics, the most efficient conditioning approach among those that result in correct type I error rates for all test statistics is used. For test statistics that simply count the number of a particular marker allele carried by affected individuals, the approach presented here in many cases, reproduces one or the other previously proposed method. In cases where more than one such method is applicable, the approach presented here provides a tool for discriminating between the methods. The approach also fills in the gaps where no previously proposed methods are applicable.

By decoupling the question of the test statistic from the problem of adjusting for admixture, the practitioner is free to use any association statistic that appears appropriate. This immediately extends family-based association statistics to analyses that combine information from extended pedigrees, applications to quantitative, survival and multivariate traits, incorporation of environmental and genetic covariates and analysis of polymorphic markers. Moreover, the practitioner is free from conceptualizing test statistics as counting transmissions from heterozygous parents. The choice of optimal test statistics requires some knowledge of the particular genetic model that is

followed by the trait and marker locus in question. However, valid inference results from the conditioning approach whether or not any assumptions that might have guided the choice of the test statistic are correct.

Finally, the exposition here has not tried to make a distinction between tests of association and tests of linkage. Instead, the goal of the analyses has been framed as either examining association between marker alleles and traits in order to detect linkage between the marker locus and a trait locus, or examining association between a candidate locus and traits with hopes of inferring a direct influence of the candidate locus on the trait. A distinction between two kinds of null hypotheses, each corresponding to a different kind of setting, has been maintained, however. One setting is where a candidate locus in a region that would not otherwise be suspected being linked to a trait locus is under examination, or where association methods are being used in a genome scan. The other setting is one where linkage is suspected in a region, and it is hoped that association methods will allow the region to be narrowed, or where linkage has been established in a region, and candidate loci in the region are under examination.

Appendix

In this appendix, the three conditions that characterize the minimal sufficient statistic under the null hypothesis are derived. The distinction between the minimal sufficient statistic based on the observed data, which is to be computed, and the full data minimal sufficient statistic, which would be used if it were available, is crucial. The full data minimal sufficient statistic is, in the case of testing for linkage, the observed traits in all pedigree members and the marker alleles in the founders. In the case of testing for association in the presence of linkage, the full data minimal sufficient statistic is the observed traits, the marker alleles in the founders and the identity-by-descent relationships.

To justify the conditions, it is sufficient to show that if two different realizations of the typed marker alleles and observed traits, y and y' have the same value of the observed data minimal sufficient statistic, then, for any value of the full data minimal sufficient statistic, x , either the conditional probabilities of y and y' given x , $P(y|x)$ 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 .

First, note that no restrictions are placed on the distribution of the full data minimal sufficient statistic under either of the two null hypotheses. It follows that, under broad regularity conditions, the full data minimal sufficient statistic is complete. (Here, complete is used in its technical sense, not in the sense of the complete data; a complete sufficient statistic for a set of models is one with the property that if the expectation of a function of the statistic is identical to a constant for every model in the set, then the function is identically equal to the constant.) Next, recall that the likelihood is the observed data minimal sufficient statistic. That is, the equivalence classes giv-

en by segregating the sample space into subsets whose members have proportional likelihood ratios are the same as the equivalence classes that correspond to the minimal sufficient statistic. The likelihood of the observed data y may be written as

$$\int dP(x)P(y|x),$$

where $P(x)$ is the marginal distribution of the complete full data sufficient statistic. The integral is over all values of the full data minimal sufficient statistic that are compatible with the observed data. Therefore, y and y' correspond to the same value of the minimal sufficient statistic if and only if, for some non-zero constant c ,

$$\int dP(x)P(y|x) = c \int dP(x)P(y'|x)$$

for every model in the null hypothesis. See, for example, Cox and Hinkley [1974].

By sufficiency, $P(y|x)$ is invariant to the distributions under the null hypothesis, so that, by completeness, y and y' correspond to the same value of the minimal sufficient statistic if and only if

$$P(y|x) = cP(y'|x)$$

for all x . The result follows directly.

Acknowledgments

This work was supported in part by grants GM 5597 and MH 59532 from the National Institutes of General Medical Sciences. The authors are grateful to Terry Speed, Ken Lange and Warren Ewens for comments on a draft. The authors are very grateful to Steve Horvath for catching several errors in the tables.

References

- Boehnke M, Langefeld CD: Genetic association mapping based on discordant sib pairs: The discordant-alleles test. *Am J Hum Genet* 1998;62:950–961.
- Cleves MA, Olson JM, Jacobs KB: Exact transmission-disequilibrium tests with multiallelic markers. *Genet Epidemiol* 1997;14:337–347.
- Cox DR, Hinkley DV: *Theoretical Statistics*. New York, Halsted Press, 1974.
- Curtis D, 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.
- Ewens WJ, Spielman RS: The transmission/disequilibrium test: History subdivision, and admixture. *Am J Hum Genet* 1995;57:455–464.
- 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.
- Knapp M: The transmission/disequilibrium test and parental-genotype reconstruction: The reconstruction-combined transmission/disequilibrium test. *Am J Hum Genet* 1999;64:861–870.
- Kaplan NL, Martin ER, Weir BS: Power studies for the transmission/disequilibrium tests with multiple alleles. *Am J Hum Genet* 1997;60:691–702.
- Lazzeroni LC, Lange K: A conditional inference framework for extending the transmission/disequilibrium test. *Hum Hered* 1998;48:67–81.
- Martin ER, Kaplan NL, Weir BS: Tests for linkage and association in nuclear families. *Am J Hum Genet* 1997;61:439–448.
- Rabinowitz D: A transmission disequilibrium test for quantitative trait loci. *Hum Hered* 1997;47:342–350.
- Schaid DJ, Li H: Genotype relative-risks and association tests for nuclear families with missing parental data. *Genet Epidemiol* 1997;14:1113–1118.
- Spielman RS, Ewens WJ: A sib-ship test for linkage in the presence of association: The sib transmission/disequilibrium test. *Am J Hum Genet* 1998;62:450–458.
- Spielman RS, McGinnis RE, Ewens WJ: Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 1993;52:506–516.
- Terwilliger JD, Ott J: A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. *Hum Hered* 1992;42:337–346.