Published in final edited form as:
Genet Epidemiol. 2012 July ; 36(5): 438–450. doi:10.1002/gepi.21638.

XM: Association Testing on the X-Chromosome in Case-Control Samples with Related Individuals

Timothy Thornton[1], **Qian Zhang**[2,4], **Xiaochen Cai**[2,5], **Carole Ober**[3], and **Mary Sara McPeek**[2,3]

[1]Department of Biostatistics, University of Washington, Seattle, WA 98195 USA

[2]Department of Statistics, University of Chicago, Chicago, IL 60637 USA

[3]Department of Human Genetics, University of Chicago, Chicago, IL 60637 USA

Abstract

Genetic variants on the X-chromosome could potentially play an important role in some complex traits. However, development of methods for detecting association with X-linked markers has lagged behind that for autosomal markers. We propose methods for case-control association testing with X-chromosome markers in samples with related individuals. Our method, X_M, appropriately adjusts for both correlation among relatives and male-female allele copy number differences. Features of X_M include: (1) it is applicable to and computationally feasible for completely general combinations of family and case-control designs; (2) it allows for both unaffected controls and controls of unknown phenotype to be included in the same analysis; (3) it can incorporate phenotype information on relatives with missing genotype data; and (4) it adjusts for sex-specific trait prevalence values. We propose two other tests, X_χ and X_W, which can also be useful in certain contexts. We derive the best linear unbiased estimator of allele frequency, and its variance, for X-linked markers. In simulation studies with related individuals, we demonstrate the power and validity of the proposed methods. We apply the methods to X-chromosome association analysis of (1) asthma in a Hutterite sample and (2) alcohol dependence in the GAW 14 COGA data. In analysis (1), we demonstrate computational feasibility of X_M and the applicability of our robust variance estimator. In analysis (2), we detect significant association, after Bonferroni correction, between alcohol dependence and SNP rs979606 in the MAOA gene, where this gene has previously been found to be associated with substance abuse and antisocial behavior.

Keywords

X-linked; score test; quasi-likelihood; pedigrees; GWAS

Address for correspondence: Timothy Thornton, Department of Biostatistics, University of Washington, Campus Box 357232, Seattle, WA 98195-7232, Phone: (206) 543-8004, Fax: (206) 543-3286, tathornt@u.washington.edu.
[4]Present address: School of Medicine and Department of Biostatistics, University of Washington, Seattle, WA 98195 USA
[5]Present address: Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032 USA

Web Resources
The URLs for data presented herein are as follows:
X_M source code, http://www.stat.uchicago.edu/mcpeek/software/index.html Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/

Introduction

Genomewide association studies are routinely conducted to identify genetic variants that influence complex disorders. Despite the potential for identifying X-linked genes that influence complex traits, association methods have primarily been developed for markers on the autosomal chromosomes and significantly less attention has been given to the analysis of markers on the X-chromosome.

For case-control association testing with autosomal markers, a number of approaches have been proposed for various types of samples including unrelated samples from a single population [Sasieni, 1997], samples with related individuals from a single population [Slager and Schaid, 2001; Bourgain et al., 2004; Thornton and McPeek, 2007], unrelated samples from structured populations [Devlin and Roeder, 1999; Pritchard et al., 2000; Price et al., 2006], and samples with both pedigree and population substructure [Thornton and McPeek, 2010; Kang et al., 2010]. However, case-control association methods for autosomal markers will typically not be valid for X-chromosome analysis. Some available methods [Zheng et al., 2007; Clayton, 2008] consider X-chromosome case-control association analysis in samples of unrelated individuals from a single population. (See Loley et al. [2011] and Hickey and Bahlo [2011] for comparisons of these methods.) An earlier method [Browning et al., 2005] proposes case-control association tests with both single markers and haplotypes, with extension to the X-chromosome, where this method allows related individuals from a single population, provided that controls are not related to cases. The X-chromosome version of this method is closely connected to the X_W test, one of the tests described in **Methods**. For certain types of study designs, family-based association tests, such as the TDT [Spielman et al., 1993] and FBAT [Rabinowitz and Laird, 2000], can be used for the analysis of both autosomal and X-chromosome markers. Family-based tests, however, are generally less powerful than case-control association methods [Risch and Teng, 1998; Bacanu and Roeder, 2000; Thornton and McPeek, 2010] and are more restrictive because they typically require genotype data for family members of an affected individual. In contrast, case-control designs can allow, but do not require, genotype data for relatives of affected individuals.

We address the general problem of case-control association testing with markers on the X-chromosome in samples with related individuals from a single population, with the pedigrees assumed known. We focus on the analysis of markers from the non-pseudoautosomal regions of the X-chromosome, where there is not homology between the X and Y chromosomes. (For analysis of markers in the pseudoautosomal regions of the X and Y chromosomes, autosomal association methods can be used.)

We propose a new method, the X_M test, for association testing of X-linked markers in samples with related individuals. The X_M test can be viewed as an extension, to X-linked markers, of the M_{QLS} test [Thornton and McPeek, 2007], for association with autosomal markers. The X_M test takes into account genetic correlations among same and different sex relatives for a valid test, and also improves power by capitalizing on the property that there is enrichment for predisposing variants in affecteds with affected relatives. Some of the properties of the X_M test are that (1) it is applicable to and computationally feasible for essentially arbitrary combinations of related and unrelated individuals, including small outbred pedigrees and unrelated individuals, as well as large complex, inbred pedigrees; (2) it distinguishes between unaffected controls and controls of unknown phenotype (i.e. general population controls) and allows for both to be included in the same analysis; (3) it incorporates phenotype information on relatives who have missing genotype data at the marker being tested; and (4) it can incorporate different trait prevalence values for males and females.

For comparison, we also propose the X_χ and X_W tests, which are extensions, to X-chromosome markers, of the corrected-χ^2 and W_{QLS} tests [Bourgain et al., 2004], respectively, for autosomal markers. Furthermore, we extend the best linear unbiased estimator (BLUE) of allele frequency for autosomes [McPeek et al., 2004] to a BLUE of allele frequency for X-chromosome markers, and we give its estimated variance.

We simulate case-control samples containing both related and unrelated individuals for various multi-locus X-linked disease models, in order to assess the type I error and compare the power of the X_M, X_W, and X_χ tests. We apply X_M to identification of X-chromosome SNPs associated with alcohol dependence (MIM 103780) in a sample of moderate-size Caucasian pedigrees from the Collaborative Study of the Genetics of Alcoholism (COGA) data [Edenberg et al., 2005] of GAW 14, and we apply it to a complex Hutterite pedigree for the identification of X-chromosome SNPs associated with asthma (MIM 600807).

Methods

Some basic assumptions about the structure of the data

Suppose the case-control study consists of genotype and phenotype data on $n + m$ sampled individuals, where we allow missing data. For a given marker, assume, without loss of generality, that n of the $n + m$ individuals have non-missing genotype data at the marker, and that these individuals are indexed by $i = 1, \ldots, n$, while m individuals have missing genotype data at the marker, and they are indexed by $i = n + 1, \ldots, n + m$. Let **D** denote the phenotype data on the $n + m$ individuals, with each individual categorized as “affected”, “unaffected” or “unknown phenotype.” Here, the designation “unknown phenotype” could be used to refer to, for example, an unphenotyped individual taken from a generic control panel. Alternatively, it could refer to an individual whose phenotype has not yet become apparent. For example, if the trait under study were Alzheimer’s disease, then unaffected individuals under a certain age might be coded as “unknown phenotype.” The $n + m$ individuals can be arbitrarily related, with the pedigree(s) that specify the relationships assumed to be known. For example, the COGA data set we analyze consists of 3 and 4-generation outbred families, while the Hutterite data set we analyze consists of a single, large, inbred pedigree. Unrelated individuals can also be included in the sample, or the entire sample could consist of unrelated individuals. We analyze the data retrospectively, i.e., we condition on **D** and treat the genotype data on the n individuals as random in the analysis. The retrospective approach is appropriate, for example, with either random or phenotype-based ascertainment. In what follows, we first give a brief overview of the M_{QLS} method for case-control association testing for autosomal markers. We then describe our model for X-chromosome data and derive the BLUE of allele frequency for an X-chromosome marker. We describe the X_M test, which is our extension of the M_{QLS} method to X-chromosome markers, and we also describe the X_W and X_χ tests, which are extensions, to X-chromosome markers, of the W_{QLS} and corrected-χ^2 tests, respectively.

Overview of the M_{QLS} method for case-control association testing of autosomal markers

For simplicity of presentation, we assume that the marker to be tested is an autosomal SNP with alleles labeled “0” and “1.” The extension to multi-allelic markers is given in Appendix C of a previous work [Thornton and McPeek, 2007]. Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ be the vector of available genotypes at the marker to be tested, where $Y_i = 0, .5$, or 1, according to whether individual i has, respectively, 0, 1 or 2 copies of allele 1 at the marker. Let p represent the frequency of allele 1 in the population from which the pedigree founders are assumed to be drawn, where $0 < p < 1$. Under the null hypothesis of no association and no linkage between marker and trait, $E_0(\mathbf{Y}|\mathbf{D}) = p\mathbf{1}$, where $\mathbf{1}$ is a column vector of length n with every entry = 1. If we further assume Hardy-Weinberg equilibrium (HWE) in the population from which the pedigree founders are drawn, then $\mathrm{Var}_0(\mathbf{Y}|\mathbf{D}) = \xi^2\mathbf{\Phi}$, where $\xi^2=\frac{1}{2}p(1-p)$ and $\mathbf{\Phi}$ is the kinship matrix with (i, j)th element equal to $2\varphi_{ij}$, where φ_{ij} is the kinship coefficient between individuals i and j, and $2\varphi_{ii} = 1 + h_i$, where h_i is the inbreeding coefficient of individual i. Assuming HWE, ξ^2 can be estimated by $\widehat{\xi}_1^2=\frac{1}{2}\widehat{p}_a(1-\widehat{p}_a)$, where

$$\widehat{p}_a=(\mathbf{1}^T\mathbf{\Phi}^{-1}\mathbf{1})^{-1}\mathbf{1}^T\mathbf{\Phi}^{-1}\boldsymbol{Y} \quad (1)$$

is the BLUE [McPeek et al., 2004] of p, where the "a" in $\hat{p}_a$ is to indicate that the estimator is for autosomes. To make the approach more robust to deviations from HWE, we can remove the assumption that $\xi^2=\frac{1}{2}p(1-p)$ and simply define ξ^2 to be the null variance of the marker genotype, $\mathrm{Var}_0(Y_i|\mathbf{D})$, for an outbred individual i, which can be estimated directly [Thornton and McPeek, 2010] by $\widehat{\xi}_2^2=(n-1)^{-1}[\boldsymbol{Y}^T\mathbf{\Phi}^{-1}\boldsymbol{Y}-(\mathbf{1}^T\mathbf{\Phi}^{-1}\mathbf{1})^{-1}(\mathbf{1}^T\mathbf{\Phi}^{-1}\boldsymbol{Y})^2]$. The test statistic for the M_{QLS} method is then given by

$$M_{QLS}=\frac{(\mathbf{V}_M^T\boldsymbol{Y})^2}{\hat{\xi}^2\mathbf{V}_M^T\mathbf{\Phi}\mathbf{V}_M}, \quad (2)$$

where the estimator $\hat{\xi}^2$ in the denominator can be taken to be either $\widehat{\xi}_1^2$ or $\widehat{\xi}_2^2$ (we typically prefer the robust form, $\widehat{\xi}_2^2$). Furthermore,

$$\mathbf{V}_M=[\mathbf{I}-\mathbf{\Phi}^{-1}\mathbf{1}(\mathbf{1}^T\mathbf{\Phi}^{-1}\mathbf{1})^{-1}\mathbf{1}^T](\mathbf{R}_N+\mathbf{\Phi}^{-1}\mathbf{\Phi}_{N,M}\mathbf{R}_M), \quad (3)$$

where $R_i = 1 - k$ if individual i is affected, $-k$ if individual i is unaffected, and 0 if individual i is of unknown phenotype, where k is an estimate of the population prevalence of the trait. (For association studies with phenotype-based ascertainment, k should ideally be an externally obtained estimate, rather than the case frequency in the sample, because the case frequency in the sample will reflect ascertainment, not population prevalence.) We define N to be the set of study individuals with non-missing genotypes at the marker being tested, and M to be the set of study individuals with missing genotypes at the marker being tested. Then $\mathbf{R}_N$ is the vector of R values for individuals in N, while $\mathbf{R}_M$ is the vector of R values for individuals in M. We define $\mathbf{\Phi}_{N,M}$ to be the cross-kinship matrix between groups N and M, with $\mathbf{\Phi}_{N,M}$ having (i, j)th entry equal to twice the kinship coefficient between the ith individual in group N and the jth individual in group M.

There are several possible interpretations of the M_{QLS} statistic of equation (2). First, M_{QLS} can be derived [Thornton and McPeek, 2007] as the quasi-likelihood score test of the null hypothesis H_0: $\gamma = 0$ in the retrospective model

$$E(Y_i|\mathbf{D})=p+\sum_{j=1}^{n+m}2\varphi_{ij}R_j, \quad (4)$$

where γ represents the association parameter. This model includes an enrichment effect, i.e., individuals with multiple affected relatives are assumed to have a higher chance of carrying a causal allele than individuals without affected relatives. For outbred individuals, it can be shown [Thornton and McPeek, 2007] that this retrospective model holds, up to terms of order $o(\gamma)$, assuming any prospective, two-allele, disease model, when the effect size γ is close to 0. Second, the M_{QLS} is closely connected [Wang and McPeek, 2009] with a retrospective likelihood score test based on the following prospective model: $P(\mathbf{D}|\mathbf{Y}) = P_0(\mathbf{D})c(\mathbf{Y}, \gamma) \prod_i e^{\gamma Y_i D_i}$. Here, $D_i = 1$ if i is affected and 0 if i is unaffected, $P_0(\mathbf{D})$ is the (unspecified) null model for the joint distribution of trait values in the absence of association with the marker being tested, where arbitrary dependence among phenotypes of related individuals is allowed, and $c(\mathbf{Y}, \gamma)$ is the normalizing constant. See Wang and McPeek [2009] for more details. A third interpretation [McPeek, in press] follows from noting that the expression $\mathbf{V}_M^T\boldsymbol{Y}$ in the numerator of the M_{QLS} test statistic in equation (2) can be rewritten as

$$\mathbf{V}_M^T\boldsymbol{Y} \sum_{i \in L \cap N} R_i(Y_i - \widehat{p}_a) + \sum_{j \in L \cap M} R_j(\widehat{Y}_j - \widehat{p}_a), \quad (5)$$

where L is the set of phenotyped individuals, $\hat{p}_a$ is the BLUE given in equation (1), and, for individuals with missing genotype, $\hat{Y}_j$ is the best linear unbiased predictor (BLUP) of Y_j under the null hypothesis, given by $\hat{Y}_j = \hat{p}_a + [\mathbf{\Phi}_{M,N} \mathbf{\Phi}^{-1}(\mathbf{Y} - \hat{p}_a\mathbf{1})]_{j-n}$. (Here, the last subscript is $j - n$ because j is the individual's index in the full set of $n + m$ individuals, while $j - n$ is the individual's index among the m individuals with missing genotype.) Thus, we can interpret the M_{QLS} as involving a form of imputation of missing genotypes by their BLUPs based on genotyped relatives. The main advantage of this form of imputation is that the uncertainty in imputation and dependence in imputation across individuals is exactly taken into account in the variance that appears in the denominator of equation (2).

Association Testing with a Bi-allelic X-linked Marker

In considering how to extend the M_{QLS} method from autosomal markers to (non-pseudoautosomal) X-chromosome markers, the most obvious difference we must deal with is that for an X-chromosome marker, a male has only one allele, inherited from his mother, while a female has two, one inherited from each parent. This difference must be taken into account in our choice of alternative mean model, which will determine the form of our quasi-likelihood score test for association. This difference will also change the correlation structure among related individuals' genotypes. One effect of the change in correlation structure is that the BLUE of allele frequency takes a different form for X-chromosome markers than for autosomes. The change in correlation structure will also affect the association test statistic. These ideas are laid out in detail in the current and following subsections.

Suppose that the marker to be tested is an X-chromosome SNP with alleles labeled "0" and "1" (the extension to multi-allelic markers is given in the Appendix). Recall that we have assumed, without loss of generality, that the first n of the $n + m$ listed individuals in the study have non-missing genotype data at the marker, while the last m individuals have missing genotype data at the marker. Let $\boldsymbol{Y}_{\mathbf{X}}=(Y_1^X, \ldots, Y_n^X)^T$ be the vector of available genotypes at the marker to be tested, where, for i female, $Y_i^X=0, .5 \text{ or } 1$, according to whether individual i has, respectively, 0, 1 or 2 copies of allele 1 at the marker, while for i male, $Y_i^X=0$ or 1, according to whether individual i has, respectively, 0 or 1 copies of allele 1 at the marker. Let p represent the frequency of allele 1 in the population from which the pedigree founders are assumed to be drawn, where $0 < p < 1$. Under the null hypothesis of no association and no linkage between marker and trait, $E_0(\mathbf{Y}_\mathbf{X}|\mathbf{D}) = p\mathbf{1}$. Note that for X-chromosome markers, we have constructed Y_i^X in such a way that, under the null hypothesis, we retain the property $E_0(Y_i^X|\mathbf{D})=p$ both for i male and for i female, as in the autosomal-marker case.

Variance of Y_X under the null hypothesis

We set $\mathbf{\Sigma_X} = \mathrm{Var}_0(\mathbf{Y_X}|\mathbf{D})$, the $n \times n$ covariance matrix of $\mathbf{Y_X}$ under the null hypothesis of no association and no linkage between the given marker and the trait. We have two approaches for estimating $\mathbf{\Sigma_X}$. The first approach assumes HWE in the founders at the marker under the null hypothesis. Under this assumption, $\mathbf{\Sigma_X} = \sigma^2\mathbf{\Phi_X}$, where $\sigma^2=\frac{1}{2}p(1-p)$, and $\mathbf{\Phi_X}$ is the X-kinship matrix with (i, j)th element equal to $2\varphi_{ij}^X$, where φ_{ij}^X is the X-Chromosome kinship coefficient between individuals i and j, which is defined in the next subsection, **X-Kinship Coefficients**. Note that σ^2 can be interpreted as the null genotypic variance, $\mathrm{Var}_0(Y_i^X|\mathbf{D})$, at the marker for an outbred female or as one-half the null genotypic variance at the marker for a male.

A reasonable estimator for $\mathbf{\Sigma_X}$ under the HWE assumption is $\widehat{\sigma}_1^2\mathbf{\Phi_X}$, where

$$\widehat{\sigma}_1^2=\frac{1}{2}\widehat{p}(1-\widehat{p}) \quad (6)$$

with $\hat{p}$ equal to the best linear unbiased estimator (BLUE) of p given below in the subsection **BLUE for Estimation of Allele Frequency for X-linked Markers**. Alternatively, for a more robust estimator that relaxes the HWE assumption at the marker, we propose $\widehat{\sigma}_2^2\mathbf{\Phi_X}$ as an estimator for $\mathbf{\Sigma_X}$, where

$$\widehat{\sigma}_2^2=(n-1)^{-1}[\boldsymbol{Y}_{\mathrm{X}}{}^T\mathbf{\Phi}_{\mathrm{X}}^{-1}\boldsymbol{Y}_{\mathrm{X}}-(\mathbf{1}^T\mathbf{\Phi}_{\mathrm{X}}^{-1}\mathbf{1})^{-1}(\mathbf{1}^T\mathbf{\Phi}_{\mathrm{X}}^{-1}\boldsymbol{Y}_{\mathrm{X}})^2], \quad (7)$$

which is RSS/$(n - 1)$ for generalized regression of $\mathbf{Y_X}$ on $\mathbf{1}$, where RSS is the residual sum of squares.

X-kinship Coefficients

The X-kinship coefficient, φ_{ij}^X, between individuals i and j is defined to be the probability that, for an X-chromosome marker, an allele drawn at random from individual i and an allele drawn at random from individual j are identical by descent. If i and j are the same individual, then the two alleles are drawn with replacement. Unlike the autosomal kinship coefficient, the X-kinship coefficient for individuals i and j depends on the sexes of i and j. Table I gives the X-kinship coefficients and autosomal kinship coefficients for a few outbred relative pairs.

For i female, we define the X-inbreeding coefficient, h_i^X, to be equal to the X-kinship coefficient of i's parents. Then, for i female, we have $\varphi_{ii}^X=\frac{1}{2}+\frac{1}{2}h_i^X$. For i male, we have $\varphi_{ii}^X=1$. We choose to define $h_i^X=1$ for i male, so that the formula $\varphi_{ii}^X=\frac{1}{2}+\frac{1}{2}h_i^X$ holds for males as well as females. If i is an outbred female, then $h_i^X=0$, while for i male, $h_i^X=1$, regardless of whether or not i is outbred. The following recursive formulas [Lange et al., 1976] can be used to calculate X-chromosome kinship coefficients for individuals i and j:

1. for j male, $\varphi_{ij}^X=\varphi_{i,mj}^X$,
2. for j female, $\varphi_{ij}^X=\frac{1}{2}\varphi_{i,mj}^X+\frac{1}{2}\varphi_{i,fj}^X$,
3. for i male, $\varphi_{ii}^X=1$, and

4. for i female, $\varphi^X_{ii}=\frac{1}{2}+\frac{1}{2}\varphi_{mi,fi}$,

where the subscripts *mi*, *fi*, *mj*, and *fj* denote, respectively, the mother of i, father of i, mother of j and father of j. The boundary conditions are $\varphi^X_{ii}=\frac{1}{2}$ if i is a female founder, $\varphi^X_{ii}=1$ if i is a male founder, and $\varphi^X_{ij}=0$ if i and j are two different founders. The KinInbcoefX Software [Zhang et al., 2009] can be used to obtain X-kinship and X-inbreeding coefficients for arbitrary pedigrees, including large, complex pedigrees with loops.

BLUE for Estimation of Allele Frequency for X-linked Markers

It is useful to have a method for fast, precise, allele frequency estimation for X-chromosome markers from data that include relatives. Previous work [McPeek et al., 2004] developed the BLUE for autosomal marker allele frequency estimation in samples with related individuals. We extend this work to derive the BLUE for allele frequencies of X-linked markers in samples with related individuals. Intuitively, the BLUE is the weighted average, of the elements of the genotype vector $\mathbf{Y}_{\mathbf{X}}$, that is optimal for estimating the allele frequency p. This optimal weighted average is obtained using the fact that, under the null hypothesis of no association and no linkage between the X-chromosome marker and trait value, our model specifies $E_0(\mathbf{Y}_{\mathbf{X}}|\mathbf{D}) = p\mathbf{1}$ and $\mathrm{Var}_0(\mathbf{Y}_{\mathbf{X}}|\mathbf{D}) = \sigma^2\mathbf{\Phi}_{\mathbf{X}}$, where $\sigma^2 > 0$ is arbitrary (i.e., not necessarily $=\frac{1}{2}p(1-p)$). In that case, the BLUE for p is

$$\widehat{p}=(\mathbf{1}^T\mathbf{\Phi}_{\mathrm{x}}^{-1}\mathbf{1})^{-1}\mathbf{1}^T\mathbf{\Phi}_{\mathrm{x}}^{-1}\boldsymbol{Y}_{\mathrm{x}}, \quad (8)$$

and the variance of $\hat{p}$ is

$$\mathrm{Var}(\widehat{p})=\sigma^2\,(\mathbf{1}^T\mathbf{\Phi}_{\mathrm{x}}^{-1}\mathbf{1})^{-1} \quad (9)$$

If HWE holds at the marker, then $\sigma^2=\frac{1}{2}p(1-p)$, which is the variance for an outbred female, and $\widehat{\sigma}^2_1=\frac{1}{2}\widehat{p}(1-\widehat{p})$ can be used to estimate σ^2. For a more robust estimator of σ^2, $\widehat{\sigma}^2_2$ of equation (7) can be used.

The matrix $\mathbf{\Phi}_{\mathbf{X}}$ will be invertible except in the case when there are mono-zygotic (MZ) twins in the sample. See Supplementary Materials for the extension of all the methods of this paper to samples containing MZ twins.

General Form of Our X-chromosome Association Test Statistics

For testing association between an X-chromosome marker and a complex trait, we consider statistics of the form

$$\frac{(\mathbf{V}^T\boldsymbol{Y}_{\mathrm{x}})^2}{\widehat{\sigma}^2\mathbf{V}^T\mathbf{\Phi}_{\mathrm{x}}\mathbf{V}}, \quad (10)$$

where $\mathbf{V} \neq \mathbf{0}$ is a vector that includes phenotype and pedigree information, which we condition on in the analysis. We constrain $\mathbf{V}^T\mathbf{1} = 0$, because this will imply $E_0(\mathbf{V}^T\mathbf{Y}_X|\mathbf{D}) = 0$. Under our assumptions and the usual regularity conditions, a statistic of this form will be χ^2_1-distributed under the null hypothesis of no association and no linkage. Notice that equation (10) for X-linked markers has a similar form to that of the M_{QLS} test statistic for autosomal markers given in equation (2). In what follows, we specify 3 different choices of $\mathbf{V}$, leading to the X_M, X_W and X_χ tests. In each case, we could add a subscript 1 or 2 to the test statistic, to denote use of estimator $\widehat{\sigma}_1^2$ of equation (6) or $\widehat{\sigma}_2^2$ of equation (7), respectively, for $\hat{\sigma}^2$ in equation (10). We typically prefer the robust form, $\widehat{\sigma}_2^2$.

Direction of Deviation

One of the most common statistical methods for testing for genotype-phenotype association in a sample of unrelated cases and controls is the Armitage trend test [Sasieni, 1997], which results in a statistic that is χ_1^2-distributed under the null hypothesis of no association. One interpretation of the Armitage trend test in this context is that it specifies a direction of deviation (for the log-likelihood) from the null hypothesis of no association and then tests for deviation in the specified direction. The direction of deviation specified by the Armitage trend test is the optimal one for detecting a causal locus whose marginal effect is either additive or multiplicative, on either the penetrance or logit-penetrance scale, with effect size tending to zero, in an outbred, unrelated, case-control sample with only one type of control (either unaffected controls or controls of unknown phenotype, but not both types in the same sample). The Armitage trend test will be asymptotically optimal in that case. For most other genetic models, the test will not be optimal, but it can still have reasonably good power to the extent that the true alternative model has a strong component of deviation in the direction specified by the test. The M_{QLS} test has similar properties to the Armitage trend test in that it is asymptotically optimal for detecting a causal locus whose marginal effect is either additive or multiplicative, on either the penetrance or logit-penetrance scale, with effect size tending to zero. In addition, the M_{QLS} test extends that optimality to a wider class of samples than that covered by the Armitage trend test. Whereas the asymptotic optimality of the Armitage trend test holds for outbred, unrelated case-control samples with only one type of control, the asymptotic optimality of the M_{QLS} test holds also for case-control samples containing related individuals, with both unaffected controls and controls of unknown phenotype allowed in the sample, as well as individuals with missing genotype. As is the case for the Armitage trend test, the M_{QLS} test will not be optimal for most other models, but it can still have reasonably good power to the extent that the true alternative model has a strong component of deviation in the direction specified by the test.

The Dosage Compensation Model

In order to extend the M_{QLS} test to X-linked markers, we choose a direction of deviation for our 1-df X_M test. We would like to choose the direction of deviation of X_M to correspond to a marginal model that is additive or multiplicative on the logit-penetrance scale, with marginal effect size tending to zero, as in the Armitage trend and M_{QLS} tests. Moreover, we are required to make additional modeling choices because of the fact that the marginal model for an X-linked causal variant now involves not only a population allele frequency p and 3 possible marginal effects for the 3 female genotypes, but now also two possible marginal effects for the two male genotypes. The additional assumption that we choose to make in order to specify a direction of deviation for the X_M test is the assumption of dosage compensation. Under an assumption of dosage compensation, the genotypic effect for a male with a particular allele is equivalent to the genotypic effect for a female who is homozygous for that allele. (Other choices are possible. For instance, one could assume that males who have the type 1 allele and females who are heterozygous have the same genotypic effect, but such a model treats the alleles in an asymmetric way, which would imply some prior knowledge about how the alleles influence the trait.)

By considering test statistics of the form in equation (10), where $\mathbf{Y_X}$ is encoded in such a way that males with a particular allele have the same values as do females who are homozygous for that allele, we are implicitly assuming dosage compensation in the X_χ and X_W tests defined below, as well as in the X_M test.

It should be noted that the association tests proposed in this paper are valid regardless of the actual direction of the association. The choice of direction affects only the power, not the validity of the tests. In subsection **Power Comparison of $\mathbf{X_M}$, $\mathbf{X_W}$, and $\mathbf{X_\chi}$** of **Results**, we investigate power for settings in which the trait models follow dosage compensation as well as settings in which the dosage compensation assumption does not hold.

Sex-Specific Prevalences

In order to improve power, the M_{QLS} test makes use of information on the prevalence of a trait in the study population, when this information is available. In our new X-chromosome test, X_M, we extend this idea to improve power further by incorporating information on sex-specific prevalences, when available.

Recall that the expression for the M_{QLS} test statistic given in equation (2) is a function of the phenotypic residual vector **R**, where $R_i = 0$ if i is of unknown phenotype. If i is of known phenotype, we can re-express R_i as $R_i = 1_{\{i \text{ case}\}} - k$, where $1_{\{i \text{ case}\}}$ is the indicator function for the event that i is a case, and k is an estimate of the population prevalence. If ascertainment is population-based, then **R** can be viewed as a mean-corrected phenotype indicator. If ascertainment is phenotype-based, then **R** is also adjusted for retrospective sampling by use of a value of k estimated from external, population-based data, not from the case frequency in the sample (because case frequency in the sample would reflect ascertainment). In the X_M test, we replace **R** by **A**, where $A_i = 0$ for individuals of unknown phenotype, $A_i = 1_{\{i \text{ case}\}} - k_f$ for i female with known phenotype and $A_i = 1_{\{i \text{ case}\}} - k_m$ for i male with known phenotype, where k_f and k_m are estimates of the population prevalence of the trait in, respectively, females and males, with $0 < k_f, k_m < 1$. When ascertainment is phenotype-based, k_f and k_m should ideally be obtained from external, population-based data, rather than from the sex-specific sample case frequencies in the data set.

The test will be valid regardless of the values chosen for k_f and k_m, though accurate values can improve the power of the test. We recommend setting k_f and k_m equal to sex-specific prevalence estimates from previous studies or registry data from the population, when available. For studies with random ascertainment, k_f and k_m could be estimated from the sample. Alternatively, k_f and k_m can be set equal, if sex-specific prevalence information is not available or if there is no evidence of a sex difference in prevalence. Robustness of power of the test to the misspecification of k_f and k_m is investigated in subsection **Robustness of power of X_M to choice of k_f and k_m**.

The X_M Test

Let **A** be the phenotype vector centered by sex-specific prevalence, as defined in the previous subsection. Without loss of generality, we can write $\mathbf{A}=(\mathbf{A}_N^T, \mathbf{A}_M^T)^T$, where $\mathbf{A}_N$ is the centered phenotype vector for the n individuals with non-missing genotype data at the marker being tested, and $\mathbf{A}_M$ is the centered phenotype vector for the m individuals with missing genotype data at the marker being tested. The X_M test statistic is given by equation (10) with **V** set equal to $\mathbf{V}_{XM}$, which is defined to be

$$\mathbf{V}_{XM}=[\mathbf{I}-\mathbf{\Phi}_X^{-1}\mathbf{1}(\mathbf{1}^T\mathbf{\Phi}_X^{-1}\mathbf{1})^{-1}\mathbf{1}^T](\mathbf{A}_N+\mathbf{\Phi}_X^{-1}\mathbf{\Phi}_X^{NM}\mathbf{A}_M). \quad (11)$$

Here, $\mathbf{\Phi}_X^{NM}$ is the $n \times m$ X-kinship matrix between individuals in groups N and M, with $\mathbf{\Phi}_X^{NM}$ having (i, j)th entry equal to twice the X-kinship coefficient between the ith individual in group N and the jth individual in group M. Note that equation (11) has a similar form to equation (3), with $\mathbf{R}_N$, $\mathbf{R}_M$, $\mathbf{\Phi}$, and $\mathbf{\Phi}_{N,M}$ replaced by $\mathbf{A}_N$, $\mathbf{A}_M$, $\mathbf{\Phi}_X$, and $\mathbf{\Phi}_X^{N,M}$, respectively.

We could add subscripts 1 and 2 to the name of the test, i.e., X_{M1} and X_{M2}, to distinguish the use of estimators $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$ given in equations (6) and (7), respectively. (We recommend use of $\widehat{\sigma}_2^2$.) X_M is asymptotically χ_1^2 distributed under the null hypothesis.

As we did for the M_{QLS}, we have several justifications or interpretations for the choice of $\mathbf{V}_{XM}$ in the definition of X_M. First, the X_M statistic can be derived as the quasi-likelihood score test of the null hypothesis H_0: $\gamma = 0$ in the retrospective model

$$E(Y_i^X|\mathbf{D})=p+\gamma\sum_{j=1}^{n+m}2\varphi_{ij}^X A_j, \quad (12)$$

where we constrain γ to be sufficiently small that $E(Y_i^X|\mathbf{D})$ lies between 0 and 1 for all i. This model captures the important feature that individuals with affected relatives are likely to be enriched for the predisposing allele relative to individuals without affected relatives. The main difference between the interpretations of the X_M statistic and the interpretations of the M_{QLS} statistic given in a previous subsection is that the retrospective mean model given in equation (4) holds, up to terms of order $o(\gamma)$, for all 2-allele disease models, while the X-chromosome version given in equation (12) holds, up to terms of order $o(\gamma)$, for models that are additive or multiplicative on the logit-penetrance scale, but not for all 2-allelle disease models in general. There is also a close connection between the X_M test and the retrospective likelihood score test based on the prospective model $P(\mathbf{D}|\boldsymbol{Y}_X)=P_0(\mathbf{D})c(\boldsymbol{Y}_X,\gamma)\prod_i e^{\gamma Y_i^X D_i}$, which is described in more detail in the subsection **Overview of the M_{QLS} method for case-control association testing of autosomal markers**. A key feature of this model is that it allows for dependence among individuals' phenotypes under the null hypothesis.

Finally, we can interpret the X_M test as involving the imputation of missing genotypes by their BLUP based on relatives' genotypes, under the null hypothesis. In other words, we can write

$$\mathbf{V}_{XM}^T\boldsymbol{Y}_X=\sum_{i\in L\cap N}A_i(Y_i^X-\widehat{p})+\sum_{j\in L\cap M}A_j(\widehat{Y}_j^X-\widehat{p}), \quad (13)$$

where $\hat{p}$ is the BLUE of p and $\widehat{Y}_j^X=\widehat{p}+[\mathbf{\Phi}_X^{NM}\mathbf{\Phi}_X^{-1}(\boldsymbol{Y}_X-\widehat{p}\mathbf{1})]_{j-n}$ is the BLUP of the unobserved genotype Y_j^X.

The X_W test

We also extend the W_{QLS} association test [Bourgain et al., 2004] for autosomal markers to the X_W test for X-linked markers. In contrast to the M_{QLS} test, the W_{QLS} assumes that every individual in the sample can be classified as either case or control. In that context, let $\mathbf{1}_c$ be the case indicator, i.e. the vector of length n whose ith entry is $1_{\{i \text{ case}\}}$, which is equal to 1 if individual i is a case and 0 if individual i is a control. We define the X_W test to be the quasi-likelihood score test for H_0: $\gamma = 0$ based on the mean model $E(\mathbf{Y^X}|\mathbf{D}) = p\mathbf{1} + \gamma\mathbf{1}_c$, where this mean model can be rewritten as

$$E(Y_i^X|\mathbf{D})=\begin{cases} p+\gamma & \text{if } i \text{ is } a \text{ case} \\ p & \text{if } i \text{ is } a \text{ control}, \end{cases} \quad (14)$$

with $0 < p + \gamma < 1$. We call this statistic X_W and it has the form of equation (10) with **V** chosen to be

$$\mathbf{V}_W=\mathbf{\Phi}_X^{-1}\mathbf{1}_c-\mathbf{1}_c^T\mathbf{\Phi}_X^{-1}\mathbf{1}(\mathbf{1}^T\mathbf{\Phi}_X^{-1}\mathbf{1})^{-1}\mathbf{\Phi}_X^{-1}\mathbf{1}. \quad (15)$$

X_W is asymptotically χ_1^2 distributed under the null hypothesis.

When cases are unrelated to controls, the X_W is asymptotically equivalent to the X-chromosome version of a previously-proposed, single-marker, allelic test [Browning et al., 2005]. For a sample of outbred, unrelated individuals, X_{W_1}, the version of the X_W test that uses variance estimator $\widehat{\sigma}_1^2$, would reduce to the Z_A^2 test of Zheng et al. [2007]. However, in order to obtain robustness to deviation from HWE, we generally recommend that variance estimator $\widehat{\sigma}_2^2$ be used instead of $\widehat{\sigma}_1^2$ in all the statistics we propose. Previous work [Thornton and McPeek, 2007] for autosomal markers has shown that the W_{QLS} is generally less powerful for mapping a genetic trait than is the M_{QLS}. For X-chromosome markers, this phenomenon is demonstrated in our simulations results (see **Results**). Despite the fact that they are not powerful for mapping, the W_{QLS} and its X-chromosome version, X_W, may still be of interest because they have the potential to be useful for testing for association of genotype with a non-genetic binary trait. For example, if two samples that were genotyped under different conditions are combined in a single analysis, the W_{QLS} or X_W could be used to test for heterogenity of allele frequency between the two samples, as a quality control measure. For X-chromosome markers, this would be done by replacing the phenotype indicator, $\mathbf{1}_c$, in equation (15) by an indicator of which sample each individual was in. Further exploration of this application is beyond the scope of the present work.

The X_χ test

We extend the corrected-χ^2 association test [Bourgain et al., 2004] for autosomal markers to the X_χ test for X-chromosome markers. The X_χ is given by equation (10) with **V** chosen to be

$$\mathbf{V}_\chi=\mathbf{1}_c-\frac{n_c}{n}\mathbf{1}, \quad (16)$$

where n_c is the number of cases with available genotype data at the marker being tested for association. If $\widehat{\sigma}_1^2$ is used for $\tilde{\sigma}^2$ in the denominator of (10), the resulting statistic could be considered an extension of the 1-df Pearson χ^2 test to the X chromosome with related individuals, while if $\widehat{\sigma}_2^2$ is used for $\tilde{\sigma}^2$ in the denominator of (10), the resulting statistic could be considered an extension of the Armitage χ^2 test to the X chromosome with related individuals. In either case, X_χ is asymptotically χ_1^2 distributed under the null hypothesis.

For a sample of outbred, unrelated individuals, the X_χ test would be rather similar to the test based on U_A described in Clayton [2008], except that the genotypic variance would be estimated differently. See Supplementary Materials for details.

GAW 14 COGA data

We apply X_M to identify SNPs on the X-chromosome that are associated with alcohol dependence in data provided by the Collaborative Study of the Genetics of Alcoholism (COGA) for Genetic Analysis Workshop 14 [Edenberg et al., 2005]. There are a total of 1,614 individuals from 143 pedigrees, with each pedigree containing at least 3 affected individuals. We include in our analysis only those individuals who are coded as "white, non-Hispanic." We designate as cases those individuals who are affected with ALDX2 or who have symptoms of ALDX2, where ALDX2 is defined to be DSM-IV alcohol dependence. By these criteria, there are 447 male cases and 366 female cases with available SNP data. We designate as "unaffected controls" those individuals who are labeled as "pure unaffected," and we designate as "controls of unknown phenotype" those individuals who are labeled as "never drank alcohol." Among individuals with available SNP data, these criteria result in 35 unaffected male and 147 unaffected female controls, and 4 males and 9 females with unknown phenotype. The resulting set of 1008 individuals we analyze are from 116 pedigrees. The data set includes 310 SNPs on the X-chromosome from the Affymetrix GeneChip 10K Mapping Array.

The Hutterite Data

We analyze data from a sample of 1,415 individuals, of whom 648 are males and 767 are females, from the Hutterite population. The Hutterites are a North American religious isolate originating from Tyrol whose entire population can be traced back to 90 ancestors. The sample population has complex relatedness, with many of the individuals being related through multiple lines of descent. A more detailed description of the Hutterite population can be found in Abney et al. [2002]. The complete genealogy of the 1,415 sampled individuals was constructed from a ≥12,000-person Hutterite pedigree. This yielded a 3,673-person pedigree with 64 founders that included all known ancestors of the individuals in the sample. The sampled individuals were phenotyped for asthma. An individual was designated as a case for asthma if he or she met all 3 of the following conditions: (1) bronchial hyperresponsiveness (BHR) (≥20% drop in baseline FEV_1 at ≤25 mg/ml methacholine) or reversibility (≥15% increase in baseline FEV_1 after albuterol treatment; (2) at least two out of the following three symptoms: coughing, wheezing, shortness of breath; and 3) a doctor's diagnosis of asthma at any evaluation. Individuals who met 0 out of the above 3 conditions were designated as unaffected controls, and individuals who met 1 or 2 of the above 3 conditions were designated as controls of unknown phenotype. This phenotype definition resulted in 177 cases (83 males and 94 females), 543 unaffected controls (260 males and 283 females), and 695 controls of unknown phenotype (305 males and 390 females). Of the 1,415 individuals, 587 were genotyped on the Affymetrix GeneChip Human Mapping 500K array, 163 on the Affymetrix GeneChip Human Mapping 5.0 array, 647 on the Affymetrix GeneChip Human Mapping 6.0 array, and an additional 18 on both the 500K and 6.0 arrays. We analyze 5,826 X-chromosome SNPs that pass quality control and are present on all 3 arrays.

Results

Simulation Studies

We perform simulation studies to (1) assess the type I error of X_M, X_W, and X_χ, (2) compare the power of X_M, X_W, and X_χ, (3) assess the robustness of power of X_M to the choice of the parameters k_f and k_m, and (4) compare power for different male-female ratios among the sampled individuals. We consider two different sample configurations, each of which contains both unrelated and related cases and controls. Both sample configurations include affected and unaffected individuals from 120 outbred, three-generation pedigrees, where each pedigree has a total of 16 individuals, of whom 8 are female and 8 are male. The pattern of affected and unaffected individuals in each pedigree varies randomly according to one of the trait models described in the next subsection. Pedigrees are sampled conditional on obtaining exactly 40 pedigrees with 4 affected individuals, 40 with 5, and 40 with 6. The phenotypes of all individuals in the sampled pedigrees are observed and genotypes are observed for only a subset of individuals in each pedigree. For each individual in a sampled pedigree, the individual's genotypes are observed if and only if at least half of the individual's siblings, parents, and offspring in the sampled pedigree are affected. Sample configurations 1 and 2 both contain unrelated individuals in addition to pedigree data. The difference between sample configurations 1 and 2 lies only in the male-female ratio among the unrelated individuals in the sample. The set of unrelated individuals in sample configuration 1 comprises 50 affected females, 50 affected males, 200 unaffected females, and 200 unaffected males. The set of unrelated individuals in sample configuration 2 consists of 100 affected males and 400 unaffected males. All phenotypes and genotypes are assumed to be observed for the unrelated individuals in each sample.

Trait Models

We consider two different classes of multigene X-chromosome trait models. Model I has two unlinked causal X-chromosome SNPs with epistasis between them and both of them acting dominantly. In Model I, the frequencies of allele 1 at SNPs 1 and 2 are p_1 and p_2, respectively. Females with at least one copy of allele 1 at SNP 1 and at least one copy of allele 1 at SNP 2 have penetrance f_1. All other females have penetrance $f_2 < f_1$. Males with allele 1 at SNP 1 and allele 1 at SNP 2 have penetrance f_3. All other males have penetrance $f_4 < f_3$. We consider two different parameter settings for model I, which are listed as models I-a and I-b in Table II. Model II has two unlinked causal X-chromosome SNPs with epistasis between them, with SNP 1 acting recessively and SNP 2 acting dominantly. Females with two copies of allele 1 at SNP 1 and at least one copy of allele 1 at SNP 2 have penetrance f_1. All other females have penetrance $f_2 < f_1$. Males with allele 1 at SNP 1 and allele 1 at SNP 2 have penetrance f_3. All other males have penetrance $f_4 < f_3$. We consider two different parameter settings for model II, which are listed as models II-a and II-b in Table II. The parameter settings for each model are given in Table II. The female and male population prevelances, which we denote as k_f^* and k_m^*, respectively, can also be found in Table II for each model. (Here, the * is used to denote that these are the true prevalence values, while we use k_f and k_m to denote the (estimated or assumed) prevalence values used in calculating X_M.) Table II also gives, for each model, the expected frequencies of allele 1 at SNP 1 among female and male cases. These values can be compared to each other as well as to p_1, the population frequency of allele 1 at SNP 1, which is in the second column of Table II.

None of the four models we consider fit the assumption of a marginal effect, for the SNP being tested, that is additive or multiplicative on the logit scale. Two models (I-a and II-a) satisfy the dosage compensation assumption and the other two do not, as indicated in Table II. The male and female prevalences range from being very similar (model II-a) to being different by a factor of more than 5 (model II-b). The frequencies of the predisposing allele in male and female cases can be equal (model I-a), the frequency in female cases can be higher (I-b), or the frequency in male cases can be higher (II-b).

Assessment of Type I Error of X_M, X_W, and X_χ

We perform simulation studies under the null hypothesis of no association and no linkageto verify that the use of the χ_1^2 approximations to the null distributions of X_{M_2}, X_{W_2}, and $X_{\chi 2}$ give the appropriate type I error for the tests, where the subscript "2" denotes use of the estimator $\widehat{\sigma}_2^2$ of equation (7). The phenotype is simulated from model I-a with sample configuration 1. We test at an unlinked, unassociated, X-chromosome SNP, with three different allele-frequency settings, which are given in Table III. For each allele-frequency setting, 100,000 simulated replicates are generated and each of them is tested for association at the .0001 level, using each of the three test statistics. In Table III, for each test statistic, we report the empirical type 1 error, which we calculate as the proportion of simulations in which the test statistic exceeds the χ_1^2 quantile corresponding to nominal type 1 error level .0001. Using an exact binomial calculation, we find that the empirical type 1 error rate does not differ significantly from the nominal for any of the three statistics. Thus, the use of the χ_1^2 approximation results in an accurate assessment of significance for X_{M_2}, X_{W_2}, and $X_{\chi 2}$. In these same simulations, we also performed the X_{M_1}, X_{W_1}, and $X_{\chi 1}$ tests, which use variance estimator $\widehat{\sigma}_1^2$ of equation (6) instead of $\widehat{\sigma}_2^2$, and we obtained nearly identical empirical type 1 error rates (results not shown).

Power Comparison of X_M, X_W, and X_χ

We perform simulation studies to compare the power of X_M, X_W, and X_χ. Five thousand replicates from each of models I-a, I-b, II-a, and II-b are simulated for sample configurations 1 and 2. The test is performed at SNP 1 for each model, with the significance threshold set to .0001. Estimated power for X_{M_2}, X_{W_2}, and $X_{\chi 2}$, with standard error, is given in Table IV. In these simulations, the X_{M_2} statistic is calculated with k_f and k_m set equal to the true female and male prevalences, k_f^* and k_m^*, respectively, given in Table II.

As shown in Table IV, X_M is more powerful than both X_W, and X_χ for all but one setting. The increase in power for the X_M is substantial (a difference in power of at least .15) for model I-b and sample configuration 2, model II-a for both sample configurations, and model II-b for sample configuration 1. X_χ is slightly more powerful than X_M for model II-b with sample configuration 2. In that setting, all individuals in the unrelated set are male and the effect size of the tested variant is much larger in males than in females, representing an extreme deviation from the dosage compensation assumption. Nonetheless, X_M is nearly as powerful as X_χ in that setting. Our simulation studies indicate that X_M performs approximately as well as, or much better than, X_W and X_χ for a range of complex disease models. As in the type-I error simulation study, we also performed the X_{M_1}, X_{W_1}, and $X_{\chi 1}$ tests and obtained nearly identical power (results not shown).

Robustness of power of X_M to choice of k_f and k_m

The X_M test is valid for any choice of the parameters k_f and k_m. When k_f and k_m are equal to the population prevalences of the disease for females and males, respectively, the X_M test is asymptotically optimal for detecting association with an X-chromosome variant, under a trait model that is additive or multiplicative on the logit-penetrance scale, with dosage compensation, as the effect size tends to zero. In reality, the trait model will usually be complex, dosage compensation may not hold, and the prevalence values will be estimated. To see how the power of the X_M test is affected by different choices of k_f and k_m when dosage compensation and additivity/multiplicativity do not hold, we perform a simulation study based on the phenotype simulated from model II-b with sample configuration 1. In model II-b, the true population prevalences for females and males are $k_f^*=.212$ and $k_m^*=.041$, respectively.

Table V gives power results for the X_{M_2} test when different values of k_f and k_m are used in the analysis, where these values are chosen to be multiples of the true values. In this case, choosing k_f and k_m to be within a factor of 3 of the true population prevalences appears to give high power, suggesting that the procedure is quite robust to the choice of k_f and k_m. It is interesting that the power for X_{M_2} when k_f and k_m are set to the true prevalence values

($k_f = k_f^*$ and $k_m = k_m^*$) is slightly lower than the power when $k_f = 2k_f^*$ and $k_m = 2k_m^*$ is used. That the use of the true prevalences does not give optimal power under model II-b is reasonable. While we expect the correct prevalences to give optimal power when the causal marker has a marginal model that has dosage compensation with a small additive/multiplicative effect on the logit-penetrance scale, model II-b violates these assumptions. Even so, the power seems not too far from optimal when the true prevalences are used in model II-b.

GAW 14 COGA data

We apply X_{M_2} to test for association with alcohol dependence and 310 SNPs on the X-chromosome in the GAW 14 COGA data. The prevalences for DSM-IV alcohol dependence and abuse for White American males and White American females have been estimated [Grant et al., 2004] from the 2001–2002 National Epidemiologic Survey on Alcohol and Related Condition to be 7.5% and 2.9%, respectively. For the X_{M_2}, we accordingly set k_m = .075 and k_f = .029 in the analysis. After applying a Bonferroni correction for association testing at 310 SNPs, X_{M_2} is significant at the 5% level for SNP rs979606 ($P = 3.8 \cdot 10^{-6}$ uncorrected, .0012 corrected). SNP rs979606 is in an intron of the Monoamine oxidases A (MAOA) gene. The MAOA gene, located at Xp11.23, is a candidate gene for alcoholism. The MAOA gene is involved with the production of the enzyme monoamine oxidase. This enzyme breaks down neurotransmitters that control mood, aggression, and pleasure. Previous studies have found the MAOA gene to be associated with alcoholism [Devor et al., 1994; Nilsson et al., 2007; Nilsson et al., 2008; Ducci et al., 2008], and other substance abuse [Gade et al., 1998], as well as anti-social behavior [Manuck et al., 2000; Caspi et al., 2001; Beitchman et al., 2004].

A previous analyses of these data [Wang et al., 2011] also used the ALDX2 alcoholism phenotype and performed family-based association tests using FBAT [Rabinowitz and Laird, 2000] for individuals who self-reported as White, non-Hispanic. For SNP rs979606 in the candidate MAOA gene, the uncorrected p-value using FBAT was reported as $4.1 \cdot 10^{-4}$. The corresponding uncorrected p-value using X_M for this SNP is $3.8 \cdot 10^{-6}$, illustrating the potential improvement in power of X_M over FBAT for X-chromosome SNPs in samples with related individuals, similar to the improvement observed [Thornton and McPeek, 2007; Thornton and McPeek, 2010] for the M_{QLS} over FBAT for association testing of autosomal markers.

There were 4 additional SNPs on the X-chromosome with uncorrected p-values less than .001 using FBAT. These SNPs and their p-values were reported as rs751871 ($P = 2.3 \cdot 10^{-4}$ uncorrected), rs1591620 ($P = 4.8 \cdot 10^{-4}$ uncorrected), rs1590366 ($P = 4.1 \cdot 10^{-4}$ uncorrected), and rs742997 ($P = 9.1 \cdot 10^{-5}$ uncorrected). However, all four of these SNPs fail our quality control check because they each have a significant sex difference in allele frequency, beyond that explained by phenotype. This is apparent from performing a generalized regression of $\mathbf{Y_X}$ on $\mathbf{S}$ and $\mathbf{\Phi_X A_N}$ with intercept, where $\mathbf{S}$ is an indicator vector for sex with ith entry 0 if i is male and 1 if i is female. If we ignore the fact that they failed our quality control check, the corresponding uncorrected p-values for these SNPs by the X_{M_2} are $6.1 \cdot 10^{-10}$, $1.4 \cdot 10^{-4}$, $1.1 \cdot 10^{-3}$, and $2.1 \cdot 10^{-15}$ respectively. None of these SNPs are in genes. In contrast, SNP rs979606, in the MAOA gene, does not have a significant sex effect in the generalized regression model, and so passes our quality control check.

The Hutterite Data

We apply X_M to identify SNPs on the X-chromosome that are associated with asthma in the Hutterite data. We estimated the male and female prevalences of asthma in the Hutterites to be nearly identical (.161 for females and .166 for males), so we set both k_f and k_m in the X_M test to .16. There were no significant SNPs identified by X_M after applying a Bonferonni correction to the p-values. We analyzed the Hutterite data using both X_{M_1}, in which the variance estimator assumes HWE in the pedigree founders, and X_{M_2}, which uses the robust variance estimator, $\widehat{\sigma}_2^2$. Figure 1 gives Q-Q plots for the resulting p-values from the two statistics. In these plots, the X_{M_1} test appears to be much less well-calibrated than the X_{M_2} test, implying that the assumptions leading to variance estimator $\widehat{\sigma}_1^2$ do not appear to hold for a large number of markers. If the assumptions held, the two statistics should perform nearly identically. In general, possible explanations for this phenomenon could include (1) genotyping error, (2) additional inbreeding not accounted for by the pedigree (or just additional relatedness), (3) cryptic population structure, (4) assortative mating related to the X-chromosome, and (5) selection related to the X-chromosome. Cryptic population structure seems implausible, given the extensive, well-documented pedigree information available for the Hutterites. It is plausible that some Hutterite founders may have been related, so explanation (2) above is a possibility. Whatever the reason for the apparent deviation from assumptions, it is interesting that X_{M_2}, which involves a robust variance estimator, performs well in this sample, even when X_{M_1} does not.

Comparison of Allele Frequency Estimators

Our estimator of allele frequency for X-chromosome markers, $\hat{p}$, is the BLUE, meaning that it is the weighted average, of the elements of the genotype vector $\mathbf{Y_X}$, that is optimal for estimating the allele frequency p. To assess the performance of our estimator, we make two types of comparisons. First, for X-chromosome markers, we compare our BLUE, $\hat{p}$, to the estimator, $\overline{Y}_X$, which we call the “naive” estimator, which is the unweighted average of the elements of the genotype vector $\mathbf{Y_X}$. This comparison gives us an idea of how much efficiency is gained for X-chromosome allele frequency estimation by using the optimal weighting of the genotypes instead of equal weighting. Second, we compare our X-chromosome allele frequency estimator to autosomal allele frequency estimators, to gauge the amount of increase in uncertainty about allele frequency estimation for X-chromosomes relative to autosomes. We consider two autosomal allele frequency estimators, the BLUE for autosomes, $\hat{p}_a$ given in equation (1), and the naive estimator, $\overline{Y}$, which is the mean of the observed autosomal genotypes. To compare the estimators, we use the pedigree information on the 1008 individuals we analyzed for alcoholism in the GAW 14 COGA data. For each of the 4 estimators we consider, we can directly calculate the mean square error (MSE) as a function of allele frequency, for the given set of pedigrees. Because all 4 of these estimators are unbiased, the MSE is equal to the variance in each case. Thus, for example, the MSE of the X-chromosome BLUE, $\hat{p}$, is obtained from equation (9), and the MSE of the autosomal BLUE, $\hat{p}_a$, is similar but with $\mathbf{\Phi}_X$ replaced by $\mathbf{\Phi}$.

Figure 2 shows MSE, as a function of allele frequency, for each of the 4 estimators. Smaller MSE corresponds to better accuracy of estimation. The most striking feature of Figure 2 is how much more efficient our BLUE estimator, $\hat{p}$, is compared to the naive estimator, $\overline{Y}_X$. This shows that optimal weighting leads to much better allele frequency estimation than equal weighting. We can also see that the X-chromosome allele frequencies are estimated less precisely than autosomal allele frequencies based on the same set of individuals. This is expected because the fact that a male has only one allele generally means that the effective sample size is smaller for X-chromosome alleles than for autosomal alleles. It is interesting that in this sample, the optimal weighting seems to make a larger difference in the MSE than does the difference between X-chromosome and autosomal markers.

Assessment of Computation Time

Using a single machine with Intel Xeon quad-core 2.66 GHz processors with 16 GB RAM and calculating all three test statistics (X_W, X_χ, and X_M), analysis of 310 X-chromosome

SNPs from the COGA data took approximately 1 second, while analysis of 5,826 X-chromosome SNPs from the Hutterite data took approximately 894 minutes (14 hours and 54 minutes). For a given data set, the computation scales linearly with the number of SNPs. The large difference in per-SNP computation time between the Hutterite and COGA data sets (9.2 seconds vs. .003 seconds) is due to the large pedigree size for the Hutterite data (approximately 1,400 individuals in a single pedigree) in contrast to the much smaller pedigree sizes in COGA. This is because for each SNP we compute a matrix decomposition for each family in the data, where the dimension of the matrix is the number of sampled individuals in the pedigree for the family. This is done to allow the pattern of missing genotype data to vary across SNPs. Thus, a sample consisting of a single large pedigree is much more time-consuming to analyze than an equal-size sample consisting of many smaller pedigrees. For comparison, we also performed a similar analysis using a subset of 609 individuals from the Hutterite data, and the analysis took approximately 19 minutes, or around .2 seconds per SNP. The speed could presumably be improved, as we have not made extensive attempts to optimize the code.

Discussion

Thanks to technological advances in genome scans, genetic association studies routinely have high-density data across the genome. Despite these advances, the mapping of many complex traits has proven to be difficult, illustrating the need for new and more powerful methods to detect susceptibility loci. Association methods have primarily been developed for the analysis of markers on the autosomal chromosomes, and significantly less attention has been given to the analysis of markers on the X-chromosome, despite the potential for identifying X-linked variants that influence complex traits. We develop methods for case-control association testing of X-linked markers when some individuals in the sample are related with known kinship. Our X_M method is applicable to, and computationally feasible for, association studies with completely general combinations of family and case-control designs. The X_M appropriately adjusts for correlated alleles and genotypes among same- and different-sex relatives and also accounts for the allele copy number difference between the sexes. The method distinguishes between unaffected controls and controls of unknown phenotype and can incorporate both into the same analysis. It can also incorporate phenotype information on relatives who have missing genotype data at the marker being tested, and it uses sex-specific prevalences to improve power. The X_M can be viewed as an extension, to X-chromosome analysis, of the M_{QLS} method for autosomal analysis. We develop two other association tests for X-linked markers, the X_W and X_χ tests, which are extensions, to X-chromosome markers, of the W_{QLS} and corrected-χ^2 tests, respectively. Our simulations indicate that, while all 3 methods control type 1 error, the X_M test tends to be the most powerful of the 3 methods, over a range of multigene trait models.

In an analysis of the X-chromosome SNPs in the GAW 14 COGA data, the X_M detected significant association, after Bonferonni correction, between SNP rs979606 and alcohol dependence, in individuals who self-reported as white, non-Hispanic. SNP rs979606 is in the MAOA gene, a candidate for alcoholism. The uncorrected X_M p-value for this SNP is $3.8 \cdot 10^{-6}$, compared to a previously reported [Wang et al., 2011] uncorrected p-value of $4.1 \cdot 10^{-4}$ using FBAT.

It is useful to have a method for fast, precise, allele frequency estimation for X-chromosome markers from data that include relatives. To solve this problem we derive the BLUE of allele frequency for X-chromosome markers in samples with related individuals, along with its estimated variance. The BLUE of allele frequency is unbiased and has smaller variance than the sample allele frequency, in samples with related individuals. It is also very fast to compute, making it a practical tool for use in genome-wide association studies.

We developed a robust variance estimator, $\widehat{\sigma}_2^2$, for X-chromosome markers, which relaxes the HWE assumption. We demonstrated in the Hutterite data set that use of the robust variance estimator, instead of the variance estimator that assumes HWE, can lead to better-calibrated statistics.

Use of the X_M requires specification of the constants k_f and k_m in the test statistic. We emphasize that the test is valid for any values of k_f and k_m. To optimize power, we recommend that k_f and k_m be set to the best available estimates of the population prevalences of the trait for females and males, respectively. It is reasonable to set $k_f = k_m$ if there is no evidence of a sex difference in prevalence or if sex-specific prevalence estimates are not available. Our simulation studies suggest that the power of the test is very robust to the choice of k_f and k_m. When k_f and k_m were misspecified within a factor of 3 of the true female and male population prevalences, there was little or no loss of power in our simulations.

We have implemented the X_M test, as well as the X_W and X_χ tests, in the XM software package. The source code will be freely downloadable (see Web Resources).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was supported in part by the National Institutes of Health grants K01 CA148958 (to T.T.), R01 HG001645 (to M.S.M.), and R01 HL085197 (to C.O.). Data on alcohol dependence were provided by the Collaborative Study on the Genetics of Alcoholism (U10AA008401), through the Genetics Analysis Workshop (R01GM031575).

References

Abney M, McPeek MS, Ober C. Quantitative trait homozygosity and association mapping and empirical genome-wide significance in large complex pedigrees: fasting serum insulin level in the Hutterites. Am J Hum Genet. 2002; 70:920–934. [PubMed: 11880950]

Bacanu SA, Devlin B, Roeder K. The Power of Genomic Control. Am J Hum Genet. 2000; 66:1933–1944. [PubMed: 10801388]

Beitchman JH, Mik HM, Ehtesham S, Douglas L, Kennedy JL. MAOA and persistent, pervasive childhood aggression. Mol Psychiatry. 2004; 9:546–547. [PubMed: 15024395]

Bourgain C, Hoffjan S, Nicolae R, Newman D, Steiner L, Walker K, Reynolds R, Ober C, McPeek MS. Novel case-control test in a founder population identifies P-selectin as an atopy-susceptibility locus. Am J Hum Genet. 2004; 73:612–626. [PubMed: 12929084]

Browning SR, Briley JD, Briley LP, Chandra G, Charnecki JH, Ehm MG, Johansson KA, Jones BJ, Karter AJ, Yarnall DP, Wagner MJ. Case-Control Single-Marker and Haplotypic Association Analysis of Pedigree Data. Genet Epidemiol. 2005; 28:110–122. [PubMed: 15578751]

Caspi A, McClay J, Moffitt TE, Mill J, Martin J, Craig IW, Taylor A, Poulton R. Role of genotype in the cycle of violence in maltreated children. Science. 2002; 297:851–854. [PubMed: 12161658]

Choi Y, Wijsman EM, Weir BS. Case-control association testing in the presence of unknown relationships. Genet Epidemiol. 2009; 33:668–678. [PubMed: 19333967]

Clayton D. Testing for association on the X chromosome. Biostatistics. 2008; 9:593–600. [PubMed: 18441336]

Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999; 55:997–1004. [PubMed: 11315092]

Devor EJ, Cloninger CR, Hoffman PL, Tabakoff B. Association of monoamine oxidase (MAO) activity with alcoholism and alcoholic subtypes. Am J Med Genet. 1994; 48:209–213. [PubMed: 8135303]

Ducci F, Enoch MA, Hodkinson C, Xu K, Catena M, Robin RW, Goldman D. Interaction between a functional MAOA locus and childhood sexual abuse predicts alcoholism and antisocial personality disorder in adult women. Mol Psychiatry. 2008; 13:334–347. [PubMed: 17592478]

Edenberg HJ, Bierut LJ, Boyce P, Cao M, Cawley S, Chiles R, Doheny KF, Hansen M, Hinrichs T, Jones K, et al. Description of the data from the Collaborative Study on the Genetics of Alcoholism (COGA) and single-nucleotide polymorphism genotyping for Genetic Analysis Workshop 14. BMC Genetics. 2005; 6 (Suppl 1):S2. [PubMed: 16451628]

Gade R, Muhleman D, Blake H, MacMurray J, Johnson P, Verde R, Saucier G, Comings DE. Correlation of length of VNTR alleles at the X-linked MAOA gene and phenotypic effect in Tourette syndrome and drug abuse. Mol Psychiatry. 1998; 3:50–60. [PubMed: 9491813]

Grant BF, Dawson DA, Stinson FS, Chou SP, Dufour MC, Pickering RP. The 12-month prevalence and trends in DSM-IV alcohol abuse and dependence: United States, 1991–1992 and 2001–2002. Drug Alcohol Depend. 2004; 74:223–234. [PubMed: 15194200]

Hickey PF, Bahlo M. X chromosome association testing in genome wide association studies. Genet Epidemiol. 2011; 35:664–670. [PubMed: 21818774]

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010; 42:348–354. [PubMed: 20208533]

Lange K, Westlake J, Spence MA. Extensions to pedigree analysis. III. Variance components by the scoring method. Ann Hum Genet. 1976; 39:485–491. [PubMed: 952492]

Loley C, Ziegler A, König IR. Association tests for X-Chromosomal markers { a comparison of different test statistics. Hum Hered. 2011; 71:23–36. [PubMed: 21325864]

Manuck SB, Flory JD, Ferrell RE, Mann JJ, Muldoon MF. A regulatory polymorphism of the monoamine oxidase-A may be associated with variability in aggression, impulsiveness, and central nervous system serotonergic responsivity. Psychiatry Res. 2000; 95:9–23. [PubMed: 10904119]

McPeek MS. BLUP genotype imputation for case-control association testing with related individuals and missing data. J Comp Biol. In press.

McPeek MS, Wu X, Ober C. Best linear unbiased allele-frequency estimation in complex pedigrees. Biometrics. 2004; 60:359–367. [PubMed: 15180661]

Nilsson KW, Sjoberg RL, Wargelius HL, Leppert J, Lindstrom L, Oreland L. The monoamine oxidase A (MAO-A) gene, family function and maltreatment as predictors of destructive behaviour during male adolescent alcohol consumption. Addiction. 2007; 102:389–398. [PubMed: 17298646]

Nilsson KW, Wargelius HL, Sjoberg RL, Leppert J, Oreland L. The MAO-A gene, platelet MAO-B activity and psychosocial environment in adolescent female alcohol-related problem behaviour. Drug Alcohol Depend. 2008; 93:51–62. [PubMed: 18029114]

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38:904–909. [PubMed: 16862161]

Pritchard J, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000; 155:945–959. [PubMed: 10835412]

Rabinowitz D, Laird N. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Hum Hered. 2000; 50:211–223. [PubMed: 10782012]

Risch N, Teng J. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. Genome Research. 1998; 8:1273–1288. [PubMed: 9872982]

Sasieni P. From genotypes to genes: doubling the sample size. Biometrics. 1997; 5:1254–1261.

Slager SL, Schaid D. Evaluation of candidate genes in case-control studies, a statistical method to account for related subjects. Am J Hum Genet. 2001; 68:1457–1462. [PubMed: 11353403]

Spielman R, McGinnis R, Ewens W. Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet. 1993; 52:506–516. [PubMed: 8447318]

Thornton T, McPeek MS. Case-control association testing with related individuals: a more powerful quasi-likelihood score test. Am J Hum Genet. 2007; 81:321–337. [PubMed: 17668381]

Thornton T, McPeek MS. ROADTRIPS: Case-control association testing with partially or completely unknown population and pedigree structure. Am J Hum Genet. 2010; 86:172–184. [PubMed: 20137780]

Wang Z, McPeek MS. An incomplete-data quasi-likelihood approach to haplotype-based genetic association studies on related individuals. J Am Stat Assn. 2009; 104:1251–1260.

Wang KS, Liu X, Aragam N, Jian X, Mullersman JE, Liu Y, Pan Y. Family-based association analysis of alcohol dependence in the COGA sample and replication in the Australian twin-family study. J Neural Transm. 2011; 118:1293–1299. [PubMed: 21445666]

Zheng G, Joo J, Zhang C, Geller NL. Testing association for markers on the X chromosome. Genet Epidemiol. 2007; 31:834–843. [PubMed: 17549761]

Zhang, Q.; Bourgain, C.; McPeek, MS. KinInbcoefX: Calculation of X-chromosome kinship coefficients (computer program). Department of Statistics, University of Chicago; 2009. http://www.stat.uchicago.edu/~mcpeek/software/KinInbcoefX/index.htm

Appendix

Extension of X_M to Multi-allelic Case

We extend the X_M procedure to test for association between a trait and a multi-allelic X-linked marker. Suppose there are a allelic types at the marker, and let $\boldsymbol{Y}_{\mathbf{X}}=(\boldsymbol{Y}_{\mathbf{1}}^{\mathbf{X}},\ldots,\boldsymbol{Y}_{\mathbf{a-1}}^{\mathbf{X}})^T$ be an $[(a-1)n]$ vector where $\boldsymbol{Y}_{\mathbf{i}}^{\mathbf{X}}=(Y_{i1}^X,\ldots,Y_{in}^X)$ and

$$Y_{ij}^X=\begin{cases}\frac{1}{2}\times(\text{the number of alleles of type } i \text{ in individual } j) & \text{if } j \text{ is } a \text{ female}\\ I\{\text{the allele in individual } j \text{ is of type } i\} & \text{if } j \text{ is } a \text{ male}\end{cases}.$$

Let $\mathbf{p}=(p_1,\ldots,p_{a-1})^T$ denote the allele frequency distribution, at the marker, in the general population, where $p_i > 0$ is the frequency of allelic type i, and $\sum_{i=1}^{a-1}p_i<1$. Define $\mathbf{r}=(r_1,\ldots,r_{a-1})^T$ to be the $(a-1)$ vector of expected deviations in allele frequencies for a case randomly sampled from the population. Then the mean model for the X_M in the multi-allelic case is $E(\boldsymbol{Y}_{\mathbf{X}})=\mu=\mathbf{p}\otimes\mathbf{1}+\mathbf{r}\otimes(\mathbf{\Phi}_{\mathbf{X}}^{N,N\cup M}\mathbf{A})$, where ⊗ is the Kronecker product, and $\mathbf{1}$ is a vector of 1's of length n, i.e.

$$EY_{ij}^X=\mu_{ij}=p_i+r_i\sum_{k=1}^{n+m}2\varphi_{j,k}^X A_k,$$

where we constrain

$$0<p_i+r_i\sum_{k=1}^{n+m}\varphi_{j,k}^X A_k<1$$

for all $1 \le i \le a-1$, $1 \le j \le n$.

Under the null hypothesis of no association between the marker and the trait, we have that **r** = **0**, where **0** is a zero vector of length (a − 1). We denote by $\mathrm{Var}_0(\mathbf{Y_X}) = \mathbf{\Sigma_X}$ the $[(a-1)n] \times [(a-1)n]$ covariance matrix of $\mathbf{Y_X}$ under the null hypothesis of no association and no linkage between the given marker and the trait. If Hardy-Weinberg equilibrium (HWE) is assumed in the founders at the marker, then $\mathbf{\Sigma_X} = \mathbf{F} \otimes \mathbf{\Phi_X}$ where **F** is an $(a-1) \times (a-1)$ matrix with (i, j)th entry $\mathbf{F}_{ij}=\frac{1}{2}p_i(1-p_i)$ if $i = j$ and $\mathbf{F}_{ij}=-\frac{1}{2}p_ip_j$ if $i \neq j$. (Alternatively, we could relax the HWE assumption, as we did for the bi-allelic case, and estimate $\mathbf{\Sigma_X}$ using a robust estimator for **F**.) The X_M test for the multi-allelic case is

$$X_M=\sum_{k=1}^{a-1}\sum_{i=1}^{a-1}(\widehat{\mathbf{F}}^{-1})_{ik}(\boldsymbol{Y}_{\mathbf{k}}^{\mathbf{X}}-\widehat{\mu}_{0k})^T\alpha\mathbf{\Gamma}^{-1}\alpha^T(\boldsymbol{Y}_{\mathbf{i}}^{\mathbf{X}}-\widehat{\mu}_{0i}), \quad (17)$$

where $\alpha=\mathbf{A}_N+\mathbf{\Phi}_{\mathbf{X}}^{-1}\mathbf{\Phi}_{\mathbf{X}}^{N,M}\mathbf{A}_M$,

$$\mathbf{\Gamma}=\alpha^T(\mathbf{\Phi}_{\mathbf{X}}\mathbf{A}_N+\mathbf{\Phi}_{\mathbf{X}}^{N,M}\mathbf{A}_M)-(\mathbf{1}^T\alpha)^2(\mathbf{1}^T\mathbf{\Phi}_{\mathbf{X}}^{-1}\mathbf{1})^{-1},$$

$\widehat{\mu}_{0i}^{X}=\widehat{p}_{0i}\mathbf{1}$ for each i, $\hat{\mathbf{p}}_0 = (\hat{p}_{01}, \ldots, \hat{p}_{0a-1})^T$ is the maximum quasi-likelihood estimate of **p** when **r** = **0**, or equivalently the BLUE of **p**, which can be shown to be

$\widehat{p}_{0i}=(\mathbf{1}^T\mathbf{\Phi}_{\mathbf{X}}^{-1}\mathbf{1})^{-1}\mathbf{1}^T\mathbf{\Phi}_{\mathbf{X}}^{-1}\boldsymbol{Y}_{\mathbf{i}}^{\mathbf{X}}$ for each i. If HWE is assumed in the founders, then $(\mathbf{F}^{-1})_{ik}$ is the (i, k) entry of $\mathbf{F}^{-1}$ evaluated at $\hat{\mathbf{p}}_0$. (Alternatively, a more robust estimator for **F**, that relaxes the HWE assumption, could be used.) Under the null hypothesis, the X_M statistic follows a χ^2 distribution with a − 1 degrees of freedom.

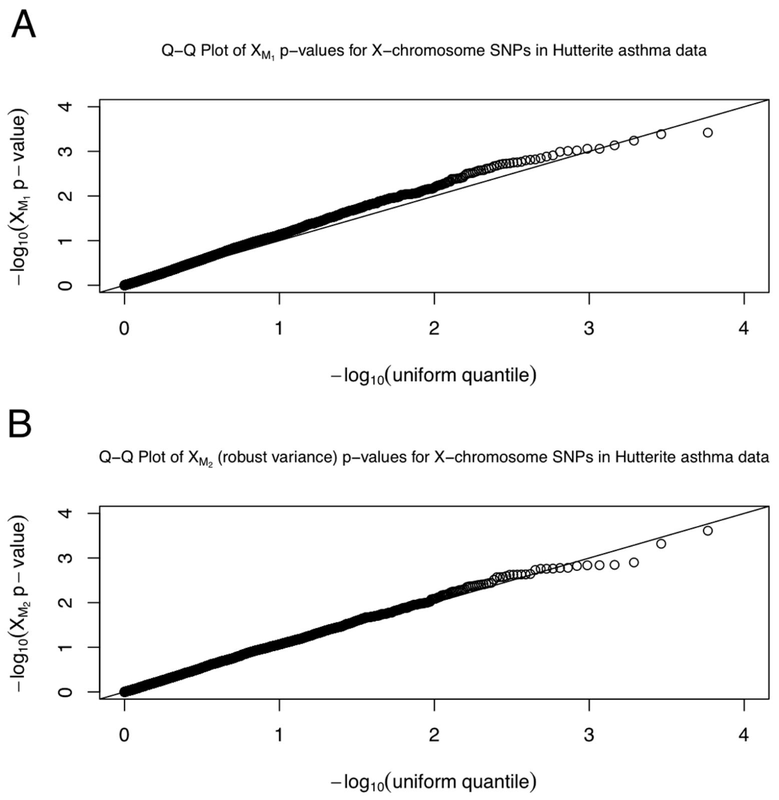

Figure 1. Q-Q plots of p-values from X_M for the Hutterite asthma data

Q-Q plots of p-values from X_{M_1} (panel A) and X_{M_2} (panel B), for 5,826 X-chromosome SNPs in Hutterite asthma data, plotted on $-\log_{10}$ scale.

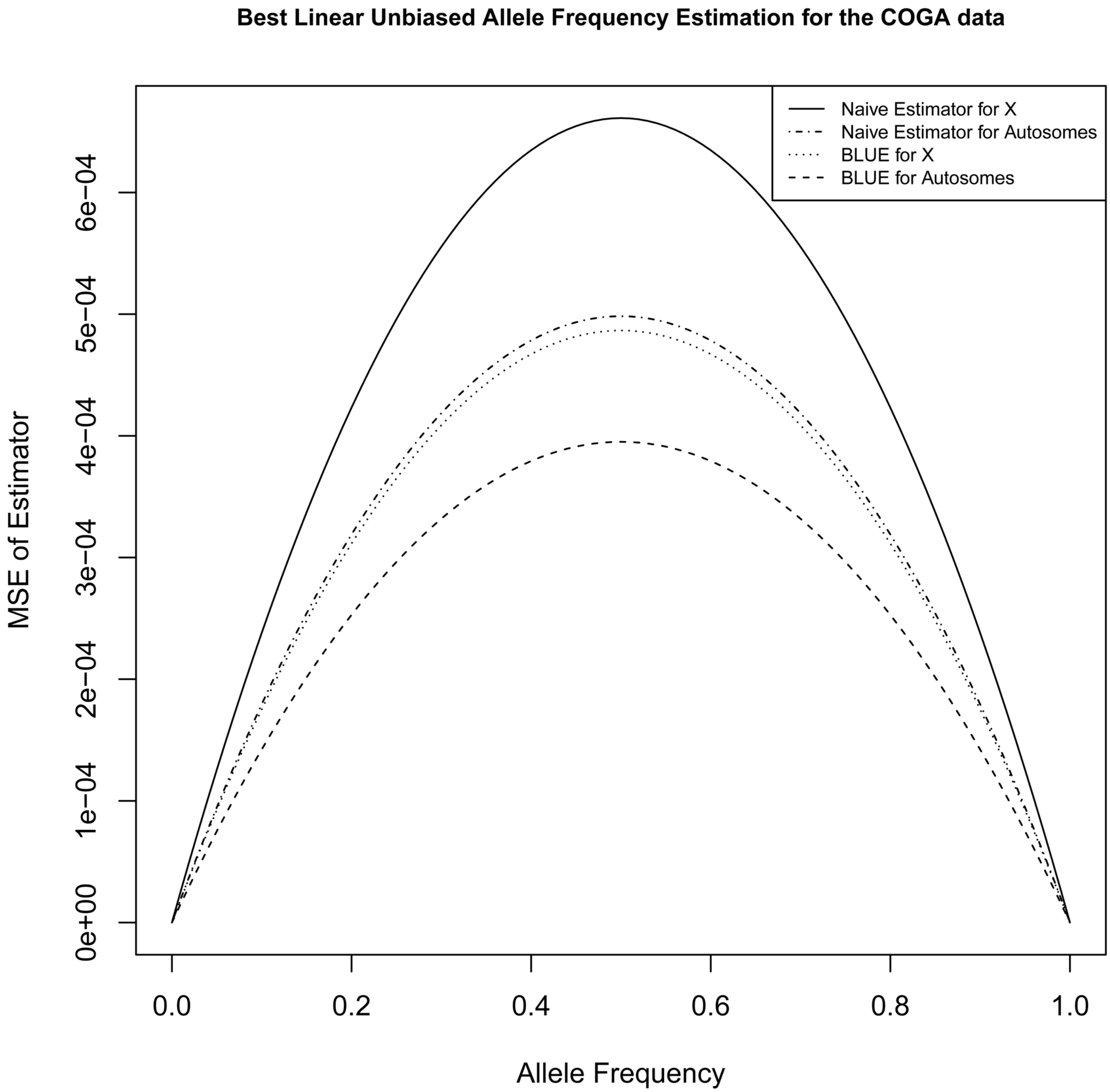

Figure 2. Mean square error (MSE) versus allele frequency in COGA pedigrees

Smaller MSE corresponds to better accuracy. Naive Estimator for X is $\bar{Y}_X$. Naive Estimator for Autosomes is $\bar{Y}$. BLUE for X is given by equation (8). BLUE for Autosomes is given by equation (1).

Table I

Examples of Autosomal and X-Kinship Coefficients for Outbred Nuclear Family Relationships

Relationship of *j* to *i*	$\boldsymbol{\varphi_{ij}}$	φ_{ij}^{X}
Mother of *i*	$\frac{1}{4}$	$\frac{1}{4}$ if *i* is female, $\frac{1}{2}$ if *i* is male
Father of *i*	$\frac{1}{4}$	$\frac{1}{2}$ if *i* is female, 0 if *i* is male
Full sister of *i*	$\frac{1}{4}$	$\frac{3}{8}$ if *i* is female, $\frac{1}{4}$ if *i* is male
Full brother of *i*	$\frac{1}{4}$	$\frac{1}{4}$ if *i* is female, $\frac{1}{2}$ if *i* is male
i (Self-kinship)	$\frac{1}{2}$	$\frac{1}{2}$ if *i* is female, 1 if *i* is male

Note. — In the table, φ_{ij} is the autosomal kinship coefficient for *i* and *j*, and φ_{ij}^{X} is the X-chromosome kinship coefficient for *i* and *j*.

Table II

Allele Frequencies and Penetrance Parameters for Simulation Models

	Allele Frequencies			Penetrance Parameters				Population Prevalences		SNP 1 Case Frequencies	
Model	p_1	p_2	**DC**	f_1	f_2	f_3	f_4	k_f^*	k_m^*	$p_{1_f}^{ca}$	$p_{1_m}^{ca}$
I-a	.2	.5	yes	.25	.1	.25	.1	.141	.115	.30	.30
I-b	.1	.45	no	.2	.07	.3	.14	.087	.147	.18	.14
II-a	.4	.15	yes	.15	.05	.15	.05	.054	.056	.45	.46
II-b	.3	.4	no	.4	.2	.12	.03	.212	.041	.34	.49

Note. — In the table, I and II denote two different classes of models, described in the text, while a and b refer to the particular parameter settings for each model, given in the table. p_i is the frequency of allele 1 at SNP i. DC is an abbreviation for dosage compensation. Penetrance parameters f_1, f_2, f_3 and f_4 are defined in the text for model classes I and II. k_f^* and k_m^* are the female and male population prevalences of the trait, respectively. p_{1f}^{ca} and p_{1m}^{ca} are the frequencies of allele 1 at SNP 1 in female cases and male cases, respectively.

Table III

Empirical Type I Error at Level .0001

	Empirical Type I Error of Tests		
p	$\boldsymbol{X_{M_2}}$	$\boldsymbol{X_{W_2}}$	$\boldsymbol{X_{\chi 2}}$
.4	.00006	.00008	.00014
.2	.00009	.00005	.00009
.05	.00010	.00016	.00015

Note. — Empirical type I error rates are calculated based on 100,000 simulated replicates. Simulation model is I-a with sample configuration 1. Association is tested with a bi-allelic X-linked marker having minor allele frequency *p*. Using an exact binomial calculation, we determined that empirical type I error rates between .00004 and .00016 are not significantly different from the nominal .0001 level. All values in the table fall within this range, so none are significantly different from the nominal.

Table IV

Empirical Power and SE at Level .0001

		Estimated Power (SE)		
Model	**Sample Configuration**	X_{M_2}	X_{W_2}	$X_{\chi 2}$
I-a	1	**.97 (.003)**	.92 (.005)	.87 (.005)
I-a	2	**.94 (.003)**	.84 (.005)	.79 (.006)
I-b	1	**.83 (.006)**	.76 (.006)	.60 (.007)
I-b	2	**.75 (.006)**	.54 (.007)	.39 (.007)
II-a	1	**.52 (.007)**	.18 (.005)	.14 (.005)
II-a	2	**.40 (.007)**	.17 (.005)	.12 (.005)
II-b	1	**.72 (.006)**	.47 (.007)	.57 (.007)
II-b	2	.74 (.006)	.70 (.007)	**.80 (.006)**

Note. — Power is assessed at significance level .0001, on the basis of 5,000 simulated replicates. The highest power for each simulation setting is in bold.

Table V

Robustness of X_M to Misspecification of k_f and k_m

Assumed k_f and k_m	Multiple of true k_f^* and k_m^*	Estimated Power (s.e.)
.053 and .010	$\frac{1}{4}$	.67 (.006)
.071 and .014	$\frac{1}{3}$	.66 (.006)
.106 and .020	$\frac{1}{2}$	.69 (.006)
.212 and .041	1	.72 (.006)
.423 and .082	2	.75 (.006)
.635 and .122	3	.71 (.006)
.846 and .163	4	.43 (.006)

Note. — Simulation model is II-b, with sample configuration 1. Power is assessed at significance level .0001, on the basis of 5,000 simulated replicates. k_f^* and k_m^* are the female and male population prevalences of the trait, while k_f and k_m are the values used in the calculation of X_M. The $\widehat{\sigma}_2^2$ of equation (7) is used for the calculation of X_M.