Testing Genetic Linkage with Relative Pairs and Covariates by Quasi-Likelihood Score Statistics

Daniel J. Schaid^a Jason P. Sinnwell^a Stephen N. Thibodeau^b

^aDivision of Biostatistics, Department of Health Sciences Research, and ^bDepartment of Laboratory Medicine and Pathology, Mayo Clinic College of Medicine, Rochester, Minn., USA

Key Words

Complex trait • Covariate • Gene-environment • Interaction • Heterogeneity • Linkage • Regression

Abstract

Background/Aims: Genetic linkage analysis of common diseases is complicated by the heterogeneity of genetic and environmental factors that increase disease risk, and possibly interactions among them. Most linkage methods that account for covariates are restricted to sib pairs, with the exception of the conditional logistic regression model [1] implemented in LODPAL in the S.A.G.E. software [2]. Although this model can be applied to arbitrary pedigrees, at times it can be difficult to maximize the likelihood due to model constraints, and it does not account for the dependence among the different types of relative pairs in a pedigree. **Methods:** To overcome these limitations, we developed a new approach based on score statistics for quasi-likelihoods, implemented as weighted least squares. Our methods can be used to test three different hypotheses: (1) a test for linkage without covariates; (2) a test for linkage with covariates, and (3) a test for effects of covariates on identity by descent sharing (i.e., heterogeneity). Furthermore, our methods are robust because they account for the dependence among different relative pairs within a pedigree. **Results and Conclusion:** Although application of our methods

to a prostate cancer linkage study did not find any critical covariates in our data, the results illustrate the utility and interpretation of our methods, and suggest, nonetheless, that our methods will be useful for a broad range of genetic linkage heterogeneity analyses.

Copyright © 2007 S. Karger AG, Basel

Introduction

Genetic linkage has been widely used to screen for susceptibility genes for complex human traits. Despite successes for many diseases that follow simple Mendelian transmission, there have been few successes of linkage analyses for common diseases. Common diseases are influenced by a spectrum of forces that include genetic, environmental, and behavioral risk factors. The definition of the phenotype, age-at-onset of disease, secular trends in diagnosis, and environmental risk factors that influence the phenotype, and possibly interact with the underlying genotype, can dramatically influence the magnitude of the observed linkage signal. For many complex diseases, the linkage signals have been weak to moderate, and some of the most promising findings have been difficult to replicate. A significant difficulty in most linkage studies of common diseases is controlling the many factors that can influence the linkage results. Accounting for

these factors can be vital to the discovery and understanding of susceptibility alleles for common diseases. To achieve this goal, we develop new regression methods for relative-pair linkage analyses.

To account for pair-specific covariates in affected sib pair (ASP) studies, several authors have extended the popular model-free maximum lod score (MLS) method of Risch [3] to allow the identical by descent (IBD) sharing probabilities for an ASP to depend on pair-specific covariates in non-linear regression models [4–10]. For a comparison of several methods, see Zhang et al. [11]. It is important to recognize, however, that when the regression effect is not large, the proposed non-linear regression models are not that different from the linear regression of the proportion of alleles shared IBD, π , on pair-specific covariates [12]. Gauderman and Siegmund showed that this type of linear regression can increase the power to detect linkage when gene-environment interaction is at least moderate, and give further discussion of the advantages of linear regression in the presence of gene-environment interaction [13]. Peng et al. [14] recently derived score statistics to test for linkage in the presence of gene-covariate interaction, and one of their score statistics for independent ASPs can be viewed as a score statistic resulting from a linear regression model. Because linear regression can be easier to fit, it offers a viable alternative to the non-linear approach. One needs to be cautious, however, about the assumption of homoscedasticity. For the regression of π on pair-specific covariates, homoscedasticity means that the variance of π , conditional on the covariates, does not depend on the values of the covariates. Clearly this is not valid, because π is a proportion, and if its mean depends on covariates, then so too does its variance. To account for this, one could use weighted linear regression with weights inversely proportional to the variance.

Based on previous developments of non-linear regression models for the influence of covariates on IBD probabilities [see references in 11 and 12], we develop new quasi-likelihood score statistics, which can be viewed as weighted least squares. Hence, they implicitly account for heteroscedasticity. A significant advantage of our approach is that it provides a method to account for correlations among multiple relative pairs from the same pedigree. These methods are developed to test three types of hypotheses: (1) a test for linkage without covariates; (2) a test for linkage with covariates, and (3) a test for covariate effects on IBD sharing. Application of these methods to a prostate cancer linkage study illustrates their utility and offers guidance for their interpretation.

Methods

Quasi-Likelihood Score Function

Our goal is to derive a flexible method to account for pair-specific covariates in linkage analyses, and to account for correlations induced by using multiple pairs of relatives from the same pedigree. Because an exact likelihood for linkage with covariates can be complex, particularly when considering the dependence among multiple pairs of relatives, we base our methods on a quasi-likelihood (QL) score function [15, 16], as others have done for complex pedigree analyses [17].

To develop a QL score function, we first consider how the expected allele sharing for a pair of relatives depends on their affection status and their covariates. First, let s_r denote the number of alleles shared IBD for a pair of relatives with relationship of type r . Because incomplete linkage information is common, we use the imputed value, $s_r = 2\hat{f}_{r,2} + \hat{f}_{r,1}$, where $\hat{f}_{r,2}$ and $\hat{f}_{r,1}$ are the estimated posterior probabilities of sharing 2 and 1 alleles IBD, respectively, conditional on the marker data of the pedigree. Second, let m_r denote the expected value of s_r , conditional on affection status and covariates. A key aspect in the development of a QL score function is how m_r depends on the covariates for a pair of relatives. Although this will be fully developed in the next section, we first present the general setup for the QL score function. Assuming that m_r depends on a vector of regression coefficients, denoted β , with vector length q , the QL score function can be expressed as

$$U = \sum_{i=1}^n D_i^T V_i^{-1} (S_i - M_i), \quad (1)$$

where i indexes a pedigree, n is the number of pedigrees, the vector S_i contains the s_r values for all k_i pairs within the i -th pedigree, the vector M_i is the expected value of S_i , the matrix D_i , with dimension $k_i \times q$ (and D_i^T is its transpose), contains the derivatives $\partial M_i / \partial \beta$, and V_i is the covariance matrix for the S_i vector, with dimension $k_i \times k_i$. To test hypotheses regarding how parameters influence the vector of means, M_i , test statistics can be derived that have the general form $T = U^T V_U^{-1} U$, where U and its covariance matrix, V_U , are computed under the null hypothesis. We use this general approach to derive a variety of score statistics, some of which are constrained to test one-sided alternative hypotheses that are in the direction of excess allele sharing (i.e., which favor linkage).

Model for Influence of Covariates on Expected Allele Sharing

To consider how linkage and covariates influence the mean function m_r , note that when there is no linkage, the expected value of s_r is $m_r^0 = 2\hat{f}_{r,2} + \hat{f}_{r,1}$, where $\hat{f}_{r,2}$ and $\hat{f}_{r,1}$ denote the no-linkage prior probabilities of sharing 2 and 1 alleles IBD, respectively. The prior probabilities depend only on relationships, not the marker data. For example, for sib pairs, $\hat{f}_{r,2} = 1/4$ and $\hat{f}_{r,1} = 1/2$. When there is linkage without covariate effects, the expected value of s_r depends on the affection status of the members of the pair, the penetrances and allele frequencies of the underlying disease susceptibility (DS) locus, and the genetic distance between the marker and the DS locus. If there is linkage with covariate effects, the expected value of s_r will additionally depend on the covariates of the relative pairs. Hence, a key aspect in our development is how we model the influence of linkage and covariates on m_r .

For now we restrict our focus to only affected relative pairs. Later we expand our model to consider relative pairs with both members unaffected and relative pairs that are discordant for affection status. For affected relative pairs, we show in Appendix 1 that the influence of pair-specific covariates on the expected allele sharing can be modeled as

$$m_r = m_r^o + c_r X \beta, \quad (2)$$

where matrix X contains the pair-specific covariates (including an intercept of 1's in the first column), and c_r is a factor that scales the covariates according to the type of relative pair. The factor c_r depends only on the prior probabilities $f_{r,1}$ and $f_{r,2}$. In Appendix 1 we derive two types of scaling factors: one that assumes no dominance variation of the penetrances, and one that approximates the minimax constraint used by Goddard and Olson [1]. Expression (2) illustrates that the expected allele sharing can be approximated by a linear regression model with an offset of m_r^o and a new covariate that is scaled according to the type of relative pair; the scaled covariate is $X^* = c_r X$. This allows us to include different types of relative pairs in the score statistic by using a scaled covariate in place of the original covariate.

If we did not scale the covariates, we would need to fit a separate regression model for each type of relative pair, requiring too many parameters. Hence, by scaling the covariates, the regression model requires fewer terms. To understand the role of scaling, consider two types of affected relative pairs from the same pedigree: an affected sib pair and an affected cousin pair. Without covariate effects, we know that the affected sib pair is expected to have greater allele sharing than the affected cousin pair, because sibs are genetically more similar than cousins. If we assumed that the intercept were the same for both types of relative pairs, we would underestimate the true intercept for the sib pair, and overestimate the true intercept for the cousin pair. By scaling the covariate according to the type of relative pair, we allow for the fact that the cousin pair is a priori expected to have less allele sharing than the sib pair, yet we require only a single intercept. Now, when covariates are added to the model, scaling the covariates achieves the same goal as scaling the intercept: the expected allele sharing for pairs can decrease as the within-pair prior genetic similarity decreases, even if the covariate values do not change. Furthermore, these methods apply to the X chromosome, noting that if a pair has at least one male, $f_2 = \hat{f}_2 = 0$.

Hypotheses for Linkage

Using the linear model for expected allele sharing (expression 2), we develop QL score statistics for three types of hypotheses:

1. test for linkage without covariates (i.e., only the intercept is of interest), in contrast to the null hypothesis of no linkage,
2. test for linkage with covariates, in contrast to the null hypothesis of no linkage,
3. test for covariate effects on IBD sharing, in contrast to the null hypothesis that the IBD sharing does not depend on covariates (e.g., linkage homogeneity), yet allowing linkage under this null hypothesis.

Test for Linkage, with or without Covariates

To develop a test for linkage with or without covariates, we consider a general model that includes an intercept, β_0 . If pair-specific covariates are included, say p covariates, there would be

a total of $q = p + 1$ regression coefficients. The test for linkage, with or without covariates, tests the null hypothesis that all regression coefficients are zero; a special case is when there is only an intercept ($q = 1$). The score vector for this hypothesis can be expressed as

$$U = \sum_{i=1}^n X_i^{*T} V_{o,i}^{-1} (S_i - M_i^o), \quad (3)$$

where the vector M_i^o contains the null expected values for the different types of affected relative pairs (m_r^o for relatives of type r), the $k_i \times q$ matrix X_i^* contains the scaled covariates, and $V_{o,i}$ is the covariance matrix for the vector S_i , computed under the null hypothesis (see Appendix 1). We later discuss how to calculate $V_{o,i}$. The covariance matrix for the vector U is

$$V_U = \sum_{i=1}^n X_i^{*T} V_{o,i}^{-1} X_i^* \quad (4)$$

and the resulting test statistic is

$$T_1 = U^T V_U^{-1} U. \quad (5)$$

For a large sample, T_1 has a χ_q^2 distribution.

One-Sided Tests Favoring Excess Allele Sharing

To increase the power to detect linkage, a variety of constraints on the IBD sharing proportions for ASP linkage studies have been proposed, ranging from Holman's 'triangle constraints', to a model with no dominance variation [18], to a model with a minimax optimality property [19]. These types of constraints have been evaluated for ASP regression models with pair-specific covariates, whereby the regression parameters were constrained such that the fitted IBD proportions met the imposed constraints. Some constraints were found to increase power when covariates were included, particularly in the presence of gene-environment interaction [6]. For instance, if an environmental exposure changes the direction of the effect of a DS allele and two affected sibs have different exposures, then their IBD sharing should be *less* than the null, while other $ASPs$ with concordant exposure should share *more* than the null. These forces, going in opposite directions from the null, lead to a large regression coefficient for the environmental covariate, giving an increase in power.

Greenwood and Bull evaluated an 'average' constraint, such that the sample mean of the model-fitted IBD sharing proportions was constrained, and a 'simultaneous boundary' constraint, such that all model-fitted IBD proportions were constrained (i.e., for all possible covariate configurations, the model-fitted IBD proportions met the assumed constraints) [6]. Although the simultaneous boundary constraint gave the best power for their simulations, it was the more difficult method to implement. Furthermore, it imposes greater constraints than the average constraint, which may not be desirable when the underlying relationships of gene(s) with environmental exposures on the risk of disease is not known; others have warned about the potential limitations of constraints in the presence of gene-environment interaction [6, 20, 21]. For these reasons, we used a simple 'average' constraint in our QL score statistics, such that the sample average of the expected number of alleles shared IBD , determined by equation (2), is not less than the sample average of the null values.

To implement the 'average constraint', all pair-specific covariates are centered about their mean values (e.g., the j -th covariate

for pair i is $(x_{i,j} - \bar{x}_j)$, these covariates are scaled, and then a score statistic is computed that considers alternative hypotheses with the intercept constrained to be non-negative. To derive this constrained score statistic, we partition the U vector of equation (3) into $U^T = (U_0 \mid U_1^T)$, where U_0 is the contribution from the intercept, and U_1 is a vector of length p for the contribution of the pair-specific covariates. In an analogous fashion, we partition the variance matrix in equation (4) into the corresponding components,

$$V_U = \begin{pmatrix} V_{00} & V_{01} \\ V_{10} & V_{11} \end{pmatrix}. \quad (6)$$

Then, by ‘zeroing-out’ the U_0 component whenever it is less than zero (i.e., average *IBD* sharing less than the null value), we compute the following constrained score statistic

$$T_{1,\text{constrained}} = \begin{cases} U^T V^{-1} U & \text{if } U_0 > 0; \text{ rank} = q \\ U^{*T} V^{*-1} U^* & \text{if } U_0 \leq 0; \text{ rank} = q-1 \end{cases}. \quad (7)$$

Here, rank is the rank of V or V^* , and U^* is the projection of U onto the half-plane $\{U_0 \geq 0\}$, which is accomplished by the regression of U_1 on U_0 . That is, $U^* = U_1 - (V_{10}/V_{00})U_0$. This regression adjusts for any correlation of the U_1 scores with the U_0 score. The variance matrix for this adjusted score vector is $V^* = V_{11} - (V_{10}V_{01})/V_{00}$. For large samples, $T_{1,\text{constrained}}$ has a chi-bar-squared distribution, which is a 50:50 mixture of χ_{q-1}^2 and χ_q^2 . These results follow from case 2 of Self and Liang [22].

It is important to note that both T_1 and $T_{1,\text{constrained}}$ can be used to test for linkage without the use of covariates, by simply including only the intercept in the score statistics. This provides a way to test for linkage using all affected relative pairs. Without covariates, the distribution of $T_{1,\text{constrained}}$ reduces to a 50:50 mixture of χ_0^2 and χ_1^2 , a well known result for testing a one-sided alternative hypothesis for excess *IBD* sharing.

Test for Effects of Covariates on *IBD* Sharing

To test for the effects of covariates on *IBD* sharing, we wish to allow for the presence of linkage by allowing a non-zero intercept, and test whether the remaining regression coefficients differ from zero. For now, we do not constrain the intercept. To develop this score statistic, we need a robust estimate of the intercept parameter. For a model with only the intercept, the X_i^* matrix reduces to a single column of 1’s multiplied by the appropriate c_i scaling factors for the types of relative pairs; denote this scaled vector C_i . From this, we can solve the following *QL* score function to obtain a robust estimate of β_0 ,

$$U_{\beta_0} = \sum_{i=1}^n C_i^T V_i^{-1} (S_i - M_i^0 - \beta_0 C_i) \equiv 0. \quad (8)$$

The solution of this equation is the weighted least squares estimate

$$\hat{\beta}_0 = \left[\sum_{i=1}^n C_i^T V_i^{-1} C_i \right]^{-1} \left[\sum_{i=1}^n C_i^T V_i^{-1} (S_i - M_i^0) \right]. \quad (9)$$

Note that the variance matrix V_i should be computed with $\hat{\beta}_0$, suggesting that one should iterate between estimating β_0 and estimating V_i . However, because it is computationally challenging to

estimate V_i , we compute V_i under the hypothesis of no linkage (all regression coefficients, including the intercept, are zero). Although this may result in a less efficient estimator of β_0 , it is still unbiased. We shall derive a *QL* score statistic that is based on the computation of V_i under the hypothesis of no linkage, and we call this a ‘model-based’ score statistic. Nonetheless, to account for a false assumption of no linkage, we shall also use a robust variance matrix, to derive a ‘robust’ score statistic.

To develop the model-based score statistic, which accounts for the additional variation from estimating β_0 , we partition the columns of the scaled X_i^* matrix into the intercept column, denoted C_i , and the remaining columns, denoted $X_{i,[-1]}^*$. Furthermore, assuming linkage, but no covariate effect, the estimated expected allele sharing is $M_i = M_i^0 + \hat{\beta}_0 C_i$. Then, the score vector can be expressed as

$$U = \left[\sum_{i=1}^n X_{i,[-1]}^{*T} V_{o,i}^{-1} (S_i - M_i) \right]. \quad (10)$$

The model-based variance matrix of U is

$$V_{U,\text{model}} = \left[\left(\sum_{i=1}^n X_{i,[-1]}^{*T} V_{o,i}^{-1} X_{i,[-1]}^* \right) - \left(\sum_{i=1}^n X_{i,[-1]}^{*T} V_{o,i}^{-1} C_i \right) \left(\sum_{i=1}^n C_i^T V_{o,i}^{-1} C_i \right)^{-1} \left(\sum_{i=1}^n C_i^T V_{o,i}^{-1} X_{i,[-1]}^* \right) \right] \quad (11)$$

and the resulting test statistic is

$$T_{2,\text{model}} = U^T \bar{V}_{U,\text{model}}^{-1} U$$

To compute the robust score statistic, we use a robust variance matrix for $\text{Var}(U)$. To achieve this, let U_i denote the contribution of the i -th pedigree to the U score vector, and partition this vector into the part that is for the intercept ($U_{i,0}$), and the part that is for the remaining covariates ($U_{i,1}$),

$$U_{i,0} = C_i^T V_{o,i}^{-1} (S_i - M_i),$$

$$U_{i,1} = X_{i,[-1]}^{*T} V_{o,i}^{-1} (S_i - M_i).$$

From these partitioned vectors, we can compute a robust variance matrix,

$$V_{U,\text{robust}} = \left[\left(\sum_{i=1}^n U_{i,1} U_{i,1}^T \right) - \left(\sum_{i=1}^n U_{i,1} U_{i,0}^T \right) \left(\sum_{i=1}^n U_{i,0} U_{i,0}^T \right)^{-1} \left(\sum_{i=1}^n U_{i,0} U_{i,1}^T \right) \right].$$

Using this robust variance, the corresponding test statistic is

$$T_{2,\text{robust}} = U^T \bar{V}_{U,\text{robust}}^{-1} U \quad (12)$$

When the covariates do not influence the *IBD* sharing, both $T_{2,\text{model}}$ and $T_{2,\text{robust}}$ have an asymptotic χ^2 distribution with $q-1$ degrees of freedom.

Now, when we constrain the intercept to be non-negative, we still estimate $\hat{\beta}_0$ according to expression (9), but if this estimate is less than zero, we set it equal to zero, and then proceed with the above steps to compute both $T_{2,\text{model}}$ and $T_{2,\text{robust}}$. For the constrained test for covariate effects, we use the unmixed χ_{q-1}^2 distribution, since in this case the null hypothesis does not include the constrained β_0 parameter.

Models to Include Unaffected Relatives

It is well known that sampling affected relatives for linkage studies of complex traits is much more powerful than sampling unaffected relatives [23]. In practice, however, it is not unusual to sample unaffected relatives in order to infer missing genotypes of parents. These unaffected relatives can be used for two purposes. One is to determine if the inclusion of unaffected relatives increases the strength of evidence for or against linkage, and the other is to evaluate the assumption of Mendelian transmission of the marker alleles in the absence of linkage. When there is linkage, relative pairs with concordant affection status are expected to have *IBD* allele sharing that exceeds the null, but the magnitude of this excess sharing depends on the penetrance of the DS locus. As the penetrance decreases, the allele sharing for concordant *unaffected* relative pairs is expected to approach the null sharing more rapidly than that for *affected* relative pairs. This is because the variability of the genotypes of unaffected subjects increases as the penetrance decreases. On the other hand, the allele sharing for pairs with discordant affection status is expected to be less than the null, with a magnitude of deficient sharing that also depends on the penetrance.

With this background, we extend our expected allele sharing model (for now ignoring covariates) by including a parameter for each of the three types of relative pairs: concordantly affected (*AA*), concordantly unaffected (*UU*), and discordant (*AU*). We achieve this by using indicator variables for each type of pair (e.g., $X_{AA} = 1$ if an *AA* pair, etc.):

$$m_r = m_r^0 + c_r \{X_{AA} \beta_{AA} + X_{UU} \beta_{UU} + X_{AU} \beta_{AU}\}. \quad (14)$$

Furthermore, based on the above background, it is reasonable to impose the following constraints when testing for linkage,

$$\begin{aligned} \beta_{AA} &\geq \beta_{UU} \\ \beta_{UU} &\geq 0 \\ \beta_{AU} &\leq 0 \end{aligned}$$

These constraints can be represented as $R\beta \leq 0$, where the inequality is coordinate-wise, $\beta' = (\beta_{AA}, \beta_{UU}, \beta_{AU})$, and

$$R = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Although more constraints might be desirable (e.g., $|\beta_{AU}| \leq \beta_{AA}$), they are not possible, because R must be of full rank. To test the null hypothesis of no linkage ($\beta = 0$) versus the alternative hypothesis of linkage under the constraint $R\beta \leq 0$, we can use the score statistic for one-sided alternatives developed by Silvapulle and Silvapulle [24]. Using the X^* matrix that results from equation (14) to compute U and V_U of equations (3) and (4), the constrained score statistic can be expressed as

$$T_3 = \tilde{\beta}' V_{\tilde{\beta}}^{-1} \tilde{\beta} - \min \{(\tilde{\beta} - b)' V_{\tilde{\beta}}^{-1} (\tilde{\beta} - b) : Rb \leq 0\}, \quad (15)$$

where $\tilde{\beta} = V_U^{-1} U$ and $V_{\tilde{\beta}} = V_U^{-1}$. The trick of using $\tilde{\beta}$ in place of the score vector U is required because we restrict the β 's (e.g., $\beta_{AA} \geq \beta_{UU}$). We could not simply restrict the corresponding U statistics, (e.g., $U_{AA} \geq U_{UU}$), because the magnitudes of the score statistics depend on the number of pairs of each type, as well as the variation of the scores. Rather, by using $\tilde{\beta}$, we implicitly account for this information. One can view $\tilde{\beta}$ as a first-order approximate es-

timate of the unconstrained regression parameters for two reasons. First, $\tilde{\beta}$ is the first updated estimate from a Newton-Raphson procedure. Second, under alternative hypotheses that are local to the null hypothesis, in the sense of $H_A : \beta = n^{-1/2} \delta$, the expected value of $\tilde{\beta}$ is β . To compute expression (15), one only needs U and V_U that are used to compute the unconstrained score test, T_1 , and to compute $\min\{\dots\}$ using a quadratic program; we used the algorithm of Wollan and Dykstra [25]. Note that if $\tilde{\beta}$ satisfies the inequality constraints, then $\min\{\dots\} = 0$, and the first term is $\tilde{\beta}' V_{\tilde{\beta}}^{-1} \tilde{\beta} = U' V_U^{-1} U$, in which case the constrained score statistic would equal the unconstrained score statistic. If, however, the constraints are not satisfied, then $\min\{\dots\} > 0$, in which case the constrained score statistic would be less than the unconstrained score statistic. The potential gain in power by using the constrained score statistic comes from a reduction in the degrees of freedom. The distribution of the unconstrained score statistic is a χ^2_3 distribution, whereas the distribution of the constrained score statistic is a mixture of χ^2 distributions, denoted $\bar{\chi}^2$ (chi-bar-squared). This mixture is

$$\bar{\chi}^2 = \sum_{k=0}^3 w_k (RV_{\tilde{\beta}} R') \chi_k^2$$

where the mixture proportions, $w_k(\cdot)$, are functions of the matrix $RV_{\tilde{\beta}} R'$, and are given in Appendix 2.

Evaluating Unexpected Allele Sharing

As emphasized by Blackwelder and Elston [23], a critical assumption for valid linkage analysis with affected relative pairs is Mendelian transmission of the marker alleles in the absence of linkage. One way to violate this assumption is misspecification of the marker allele frequencies when parents are missing genotypes. This can cause excessive *IBD* sharing even among discordant relative pairs. Other ways to violate the assumption of Mendelian transmission are misspecified pedigree relationships and inbreeding. Following the advice of Blackwelder and Elston to evaluate Mendelian transmission, we use our developed score statistics to evaluate whether unexpected allele sharing occurs for a dataset by simply comparing the constrained versus unconstrained score statistics. Excess *IBD* sharing for *AA* pairs and *UU* pairs, and deficient sharing for *AU* pairs, are all consistent with linkage. In contrast, if any of these sharing patterns are in the opposite direction, then unexpected *IBD* sharing would be suspected. In this later case, the *p* value for the unconstrained statistic would be smaller than that for the constrained statistic, suggesting violation of Mendelian transmission, or using a misspecified model, or genotype errors.

Unaffected Relatives and Covariates

To include covariates, it is sensible to allow the covariate effects to differ for each of the *AA*, *UU*, and *AU* pair types. This implies that for p covariates, a total of $3(p+1)$ regression parameters would be required. For complex traits, it is unlikely that the magnitude of the covariate effects, relative to their standard errors, would be large enough to compensate for the large number of degrees of freedom for these types of analyses. For this reason, we do not advocate routine use of unaffected subjects to test for linkage with covariates. Nonetheless, our general approach allows such modeling when deemed appropriate.

Computation of Null Covariance for Allele Sharing among Pairs of Relatives

Under the null hypothesis, the S_i scores for a pedigree can be correlated. Three illustrative examples follow. First, for a fully informative marker, it can be shown that the scores are uncorrelated for different sib pairs within the same nuclear family, even for pairs that overlap with a sib in common (although these scores are not jointly independent). Second, it can also be shown that the correlation is -1 between the scores for a grand-child paired with each of its grand-parents (i.e., grand-parents are spouses); knowing that a grand-child and grand-parent share 1 allele *IBD* implies that the same grand-child cannot share an allele *IBD* with the other grand-parent. Third, positive correlations arise between the score of first cousins and the score between their parents (i.e., when their parents are full sibs).

Depending on the size of a pedigree, we use either an exact method or a simulation method to compute the covariance matrix $V_{0,i}$ for the vector S_i under the null hypothesis of no linkage, assuming fully informative markers. For these methods, we need to consider the joint distribution of $s_{i,j}$ and $s_{i,k}$, the estimated allele sharing for relative pair j and relative pair k , both pairs from pedigree i . For small pedigrees, the exact method enumerates all possible *IBD* states for all possible pairs of relatives, along with the null probabilities of the *IBD* states. The software Merlin [26] computes this information using the options `--ibd --matrices`. For large pedigrees that cannot be analyzed by Merlin, the covariance matrix can be approximated by simulated 'gene dropping'. For each founder of a pedigree we assign a pair of integer codes for their two alleles. If there is no inbreeding, the allele codes are unique; for f founders, we assign the integers $1, 2, \dots, 2f$. These alleles are randomly transmitted throughout the pedigree according to Mendelian segregation (for either an autosome or the X chromosome). Then, for a particular configuration of transmitted alleles, we evaluate the *IBD* status for all possible pairs of relatives, allowing us to create an instance of the S_i vector under the null hypothesis of no linkage. This process is repeated a large number of times, and the sample covariance matrix of these simulated S_i vectors is used to estimate $V_{0,i}$.

Coding Covariates

The simulation results of Greenwood and Bull [6] showed that models with covariates can improve power to detect linkage in ASP studies, particularly when there is gene-covariate interaction and some ASPs are discordant for their covariates. Further work by Peng et al. [14] showed that correctly modeling gene-covariate interaction in ASP studies can lead to substantially increased power to detect linkage, in contrast to analyses that ignore the covariates. In contrast to our approach that considers how covariates affect the mean *IBD* sharing, Peng et al. derived score tests for linkage in ASP data by starting with a penetrance model for the joint effects of a genotype and a covariate on the phenotype. Despite this different starting point, one of their derived score statistics is a special case of our T_1 statistic used to test linkage with covariates. Because they restricted their analyses to ASPs, our scaling factor, c_r , for computing scaled X covariates is not required, and $V_{0,i}$ is constant, and so it drops out of the score equation in expression (3). The variance of their score vector, V_U , is the same as ours.

Overall, Peng et al. [14] found that if the influence of the gene-covariate interaction is negligible, there can be a loss in power by

Table 1. Summary of types of affected relative pairs

Type of ARP	f_0	f_1	f_2	Number of ARPs
Full-sibs	0.25	0.5	0.25	279
Half-sibs	0.5	0.5	0.0	27
Double-first-cousins	0.5625	0.375	0.0625	1
First-cousins	0.75	0.25	0.0	114
Half-first-cousins	0.875	0.125	0.0	8
Total				429

including the covariates, roughly equivalent to a 25% loss in sample size. This is caused by the additional degrees of freedom for the covariates. In contrast, there can be a substantial gain in power if the gene-covariate interaction is strong. An advantage of their approach is that it provides insight on the best ways to code covariates that are sensitive to gene-covariate interaction. Let w_1 and w_2 denote a covariate for subjects 1 and 2 of an ASP. Peng et al. proposed a three-dimensional statistic with a pair-specific covariate vector $X = (1, w_1 + w_2, w_1 w_2)$, and a similar two-dimensional vector that eliminates the term $w_1 w_2$. They found that the two-dimensional score statistic is often more powerful than the three-dimensional score statistic, consistent with conclusions by Gauderman and Siegmund [13]. These findings suggest that simple models for covariate effects, with fewer degrees of freedom, will likely provide the desired gain in power when the covariate effects are at least of moderate strength. The simple coding of a pair-specific sum will likely achieve the desired results when there is gene-covariate interaction.

Software

Our software package, *ibdreg*, uses either SPLUS or R statistical computing environments as an interface, along with linked ANSI C code for rapid computations. The calculation of *IBD* probabilities requires an external program. We use Merlin [26], and provide PERL scripts to simplify the required steps. This software can be downloaded from our web site.

Applications to Prostate Cancer Linkage

Our newly developed methods were applied to a study of linkage for familial prostate cancer conducted at the Mayo Clinic. This study included 159 pedigrees with a total of 429 ARPs, with both microsatellite and single nucleotide polymorphism (SNP) genetic markers, and so provided high information content, despite not having genotypes for the founders of the pedigrees. Details of pedigree sampling and genotyping can be found elsewhere [27, 28]. The study was approved by the Mayo Clinic Institutional Review Board.

The number of affected men per pedigree ranged 2–7, with 78 pedigrees (49%) having two affected men, 57 (36%) having three, 17 (11%) having four, and the remaining 7 (4%) having five to seven affected men. The types of relationships are summarized in table 1 according to prior probabilities of *IBD* sharing and the numbers of ARPs. Although there were five genetically different types of ARPs, the majority were either full-sibs, half-sibs, or first-

Fig. 1. Linkage results for constrained score statistic versus LODPAL for affected relative pairs with the pair-specific covariate sum of age at diagnosis.

Fig. 2. Linkage results for constrained score statistics with and without a covariate, and for covariate effect on *IBD* sharing (robust variance) for chromosome 20, using only affected relative pairs. Pair-specific covariate is the sum of age at diagnosis.

Fig. 3. Linkage results for constrained score statistics without covariates using all relatives: separate score statistics for affected pairs (AA), for unaffected pairs (UU), and for discordant pairs (AU), and a global test for all types of pairs.

cousins. Without accounting for pair-specific covariates, the most significant evidence for linkage was on chromosome 20, with a Kong and Cox [29] exponential allele sharing LOD score of 2.4 [28]. For our new analyses with covariates, we considered three pair-specific covariates: average number affected men in a pedigree, sum of ages at diagnosis, and sum of Gleason scores. Gleason score indicates prostate tumor differentiation (ranging from 2 for well-differentiated to 10 for poorly differentiated cellular architecture) and is a measure of disease aggression, with larger values having poorer prognosis.

In addition to analyses by our new score statistics, we performed regression analyses with the LODPAL software [2], which approximates the pseudo-likelihood of *IBD* sharing probabilities by a trinomial logistic regression model, as proposed by Goddard and Olson [1] (see details in Appendix A). This model imposes the minimax constraint of Whittemore and Tu [19], which is approximately halfway between a dominant and a recessive model, and requires a regression coefficient for each pair-specific covariate.

Results

The results for simultaneously testing linkage with the covariate age at diagnosis (pair-specific sum) are presented in figure 1 for our new score statistic and for LODPAL. We present results for the approximate ‘minimax’ scaling (equation A17), which were quite close to those obtained from the ‘no dominance’ scaling (equation A10; results not shown). This figure illustrates several key points. First, the results from our score statistic and from LODPAL are generally quite close. However, when they

differ, the difference can be quite large, such as for chromosome 5, where LODPAL gives extraordinary large LOD scores, most likely due to instability of the maximization of the LODPAL pseudo-likelihood. Other examples are chromosomes 6 and 20, for which the LODPAL peak appears much too narrow for the number of meioses in these data. We interpret these differences as difficulties with maximization of the LODPAL pseudo-likelihood. Score statistics circumvent this problem by not needing to maximize a complex function. These patterns of similarities and differences were found for the other covariates, number affected men in a pedigree and the pair-sum of Gleason scores (results not shown).

Given the more striking linkage signal on chromosome 20, we focused on this chromosome to illustrate additional analyses with our score statistics. Figure 2 illustrates the three types of tests: linkage without covariates, linkage with covariates, and the covariate effect on *IBD* sharing. Although linkage with the covariate (pair-sum of age at diagnosis) slightly increased the linkage signal, the signal without the covariate was almost as large. This suggests that the covariate makes only a small contribution to the linkage signal, which is further emphasized by the broken line in figure 2 for the effect of the covariate on *IBD* sharing (which used the robust variance, not the model-based variance).

The analyses presented in figures 1 and 2 were for only affected relative pairs. To examine the influence of unaf-

Fig. 4. Linkage results for constrained versus unconstrained score statistics using only affected relative pairs.

affected relatives on linkage, we present in figure 3 the constrained score statistics for each of the AA, AU, and UU pairs, as well as the global score statistic that simultaneously considers all types of pairs, constraining each type of pair to have *IBD* sharing in the direction that favors linkage. The results in figure 3 emphasize that the unaffected subjects contribute little to linkage, and that including them diminishes the linkage signal of the global score test compared to the score test for the subset of AA pairs. This is not surprising for a complex trait like prostate cancer, where the unaffected men tend to be much less informative than affected men.

To evaluate whether there is evidence for unexpected *IBD* sharing, we used only affected men (we have few unaffected men) to compare the constrained versus unconstrained score statistics. The results, illustrated in figure 4, show that for most chromosomes the p values for the unconstrained score statistics were less extreme than those for the constrained statistics. Some notable exceptions were for chromosomes 1, 7, and 12, each of which had regions where the p values for the unconstrained statistics were more extreme than those for the constrained statistics. For more detailed diagnostic checks, we examined the pedigree-specific z-scores for these regions: $z_i = U_i/\sqrt{V_i}$, where U_i and V_i are the contributions of the i -th pedigree to the score statistic, expression (3), and its variance, expression (4). Pedigrees with extremely negative z-scores were examined for whether the relationships among pedigree members was accurate and whether the pedigrees had different ethnic backgrounds than the remaining pedigrees, which might make the marker allele frequencies questionable. Based on the z-scores, we could not find any pedigrees that demonstrated consistent unexpected sharing for all three chromosomes, nor did ethnic background for the pedigrees correlate with the large negative z-scores. Furthermore, because these discrepancies appeared for only a few chromosomes, we conclude that these differences occurred by chance. We caution, however, that blindly rerunning analyses after excluding pedigrees with large negative z-scores can falsely inflate the evidence for linkage.

Discussion

Because common diseases are influenced by both genetic and non-genetic factors, it can be crucial to account for non-genetic covariates when evaluating the strength of evidence for genetic linkage. To achieve this, we developed quasi-likelihood score statistics to test for: (1) link-

age without covariates; (2) the simultaneous effects of linkage and covariates, and (3) the effects of covariates on *IBD* sharing. Although others have developed pseudo-likelihood ratio tests for these hypotheses [1, 4, 6], our approach offers several advantages. First, score statistics are rapid to compute, allowing rapid evaluation of a wide variety of covariates. Second, score statistics avoid the numerical problems of likelihood ratio statistics that require maximizing a complex function that is likely to have multiple modes, and hence can provide odd results, particularly with sparse data. Third, unlike some other methods, ours is not restricted to only ASPs, but rather includes all types of ARPs in a single score statistic. Fourth, we provide a mechanism to scale the covariates, which allows different levels of dominance variation. Fifth, we account for the correlation among different pairs from the same pedigree. Although our methods focus on relative pairs, the unit of analysis is the pedigree. This is achieved by creating a vector, for each pedigree, of allele sharing for all possible pairs of relatives of interest. The quasi-likelihood for this vector of pair-wise allele sharing incorporates a model for how pair-wise covariates influence the allele sharing. The quasi-likelihood also uses pedigree-specific covariance matrices for the vectors of pair-wise allele sharing values. These covariance matrices offer a new approach to account for the dependence among pairs originating from the same pedigree, in contrast to methods that treat pairs of relatives as the unit of analysis. It has been reported that the performance of robust sandwich variance estimators in generalized estimating equations with small samples can be poor, with inflated Type-I errors for Wald tests and conservative Type-I errors for score tests [30]. In our context, this would occur when there are few pedigrees. Fortunately, a simple adjustment to the statistic, by multiplying by $J/(J - 1)$, where J is the number of clusters (pedigrees), seems to provide an adequate fix [30]. Alternatively, the model-based variance, calculated under the null of no linkage, might be more reliable for small sample sizes.

Although likelihood ratio statistics can have greater power than score statistics when there are large departures from the null hypothesis, this is not likely for common complex traits. Furthermore, because the likelihood methods implemented in LODPAL do not account for dependent ARPs from the same pedigree, the resulting statistics are not guaranteed to have a known distribution, even for large sample sizes. The joint dependence among ARPs can skew the computed allele sharing statistics. To overcome this problem, a variety of ad hoc weighting

schemes have been proposed. However, weighting schemes can lead to extremely conservative linkage statistics, particularly for weighted likelihood ratio tests [31]. It might, however, be possible to overcome this limitation by using second order adjustments to the likelihood ratio test, based on comparisons of the observed information matrix with an empirical estimate [32]. Nonetheless, a significant advantage of our approach is the statistically valid method of using the null covariance for the entire vector of allele sharing for all possible pairs of subjects within a pedigree.

In our implementation of score statistics, we compute both the constrained and unconstrained versions, as a way to evaluate whether the constrained version makes sense. Without covariates, the p values for the constrained statistic should be smaller than that for the unconstrained statistic. Finding opposite results should raise concerns about 'unexpected' *IBD* sharing, possibly due to misspecified marker allele frequencies, genotype errors, etc. In early stages of our developments, when using both affected and unaffected relatives, we considered using the items from the score vector, $U' = (U_{AA}, U_{UU}, U_{AU})$, as a way to examine unexpected allele sharing. For example, finding $(U_{AA}/\sqrt{V_{AA}} \ll 0$ (where V_{AA} is the diagonal of the variance matrix for the U score), would suggest unexpected sharing. However, because U_{AA} , U_{UU} , and U_{AU} are correlated, this approach was not as informative as simply comparing the results from the constrained and unconstrained analyses.

When covariates are included, Greenwood and Bull [6] caution that allele sharing patterns can fall outside the triangle constraint when there is a qualitative gene-environment interaction (implying that an exposure changes the direction of effect of the *DS* allele, such as when the gene is protective in unexposed subjects, yet confers risk in exposed subjects), and when the sample contains pairs with discordant exposure [see also 20, 21]. Because of this, they suggested that the unconstrained estimates of allele sharing should be examined before using the constrained models.

Recently, Wallace et al. [33] derived a score test for the effects of covariates on *IBD* sharing for *ASP* studies. They were motivated to evaluate whether co-phenotypes, or endophenotypes, were associated with *IBD* sharing. Their score statistic is a special case of our quasi-likelihood score statistic under the assumption of no dominance variation (see Appendix 1) and when *ASPs* are independent from each other. They also determine genome-wide statistical significance by permuting covariate vectors among the *ASPs* – a method appropriate for independent

pairs. Their method, however, does not account for the correlation of pairs from the same pedigree. In contrast, our more general approach not only allows for different types of *ARPs* and different scaling factors that allow different amounts of dominance variance, but it also accounts for the correlation among multiple *ARPs* from the same pedigree.

Although we did not compute simulation p values, a significant advantage of using score statistics is that they are rapid to compute, making simulations feasible. When doing so, it would be important to account for the correlation of information provided by multiple *ARPs* from the same pedigree. To do this in a rapid way, one could use the rapid simulation strategy for score statistics proposed by Lin [34]. He showed that under the null hypothesis, and conditional on the data, simulation p values can be computed by multiplying the observed U score statistic vectors by a standard normal random variable (which we denote r), and then computing the desired summary statistic. To account for correlations within pedigrees, one could first sum the score vectors within each pedigree, and then multiply the pedigree total score, U_i , by r . Under the null hypothesis, where $E[U] = 0$, the asymptotically equivalent statistic $\tilde{U} = \sum r_i U_i$ has the desired expectation zero and the same variance matrix V , since $r_i \sim N(0,1)$ and is independent from U_i .

The results from our applications to a linkage study of prostate cancer illustrate that the score statistics tend to be more stable than methods that require maximization of a complex function. Although this particular application suggested that the covariates did not add substantial information to the linkage analyses, the applications illustrated how our methods can be used and interpreted. Further applications, particularly to datasets that have more rich covariates or co-phenotypes, will likely prove fruitful, given the results by Wallace et al. [33] that used co-phenotypes for hypertension in a sib-pair study.

We recognize that further study of the statistical properties of our methods is warranted. Some important considerations are how to control the Type-I error rate when multiple covariates are examined in a genome-wide analysis, understanding the conditions for which including covariates provides a gain in power, and evaluating whether certain covariate scaling factors can provide high power over a range of genetic and environmental influences on linkage. These are important considerations, but require such extensive simulations, with a broad range of both genetic and covariate influences on disease susceptibility, that a separate extensive study is postponed for future work.

Web Resources

ibdreg Package: <http://mayoresearch.mayo.edu/mayo/research/biostat/schaid.cfm>.

Acknowledgement

This work was supported by the U.S. Public Health Service, National Institutes of Health, contract grant numbers GM67768, CA72818.

Appendix 1. Covariate Scaling Factors

Risch showed that the penetrances of the *DS* locus depend on the sibling risk ratio λ_s , which is the ratio of the probability that a sib of an affected person will also be affected, divided by the population disease prevalence [18, 35]. Since we work with the imputed value $\hat{f}_{r,2}$ and $\hat{f}_{r,1}$, we shall first show that their expected values, $z_{r,2}$ and $z_{r,1}$, can be expressed as functions of the sibling risk ratio λ_s and the null prior probabilities $f_{r,2}$ and $f_{r,1}$. We then express λ_s as a function of pair-specific covariates and their corresponding regression coefficients, β .

To show that $z_{r,2}$ and $z_{r,1}$ can be expressed as functions of the sibling risk ratio λ_s and the null prior probabilities $f_{r,2}$ and $f_{r,1}$, we use results from Rich [18],

$$z_{r,2} = f_{r,2} \lambda_m / \lambda_r, \quad (\text{A1})$$

$$z_{r,1} = f_{r,1} \lambda_o / \lambda_r, \quad (\text{A2})$$

where λ_m , λ_o , and λ_r are the recurrence risk ratios for a monozygotic twin (*m*), an offspring (*o*), and a general relative of type *r*, respectively. Furthermore, assuming that there is no dominance variance of the penetrances, Risch [35] showed that

$$\lambda_r - 1 = 2\phi_r V_A / K^2, \quad (\text{A3})$$

where ϕ_r is the kinship coefficient, V_A is the additive genetic variance, and K is the population disease prevalence. Note that $2\phi_r = (f_{r,2} + f_{r,1})/2 = m_r^o/2$, the latter equality following from our notation for the null expected value of s_r . Using the special case of sibs, expression (A3) implies that $(V_A/K^2) = 2(\lambda_s - 1)$. Substituting this into expression (A3) and using $2\phi_r = m_r^o/2$ allows us to express equation (A3) as

$$\lambda_r - 1 = m_r^o (\lambda_s - 1). \quad (\text{A4})$$

Now, because we assumed no dominance variance, $\lambda_m = 1 + 2(\lambda_s - 1)$ and $\lambda_o = \lambda_s$; using these terms with expression (A4) allows us to express equations (A1) and (A2) as

$$\text{and } z_{r,2} = f_{r,2} \frac{[1 + 2(\lambda_s - 1)]}{[1 + m_r^o (\lambda_s - 1)]} \quad (\text{A5})$$

$$z_{r,1} = f_{r,1} \frac{\lambda_s}{[1 + m_r^o (\lambda_s - 1)]}. \quad (\text{A6})$$

We now add and subtract $f_{r,2}$ to the right-hand side of equation (A5), and add and subtract $f_{r,1}$ to the right-hand side of equation (A6), to obtain

$$z_{r,2} = f_{r,2} + f_{r,2} \frac{(\lambda_s - 1)(2 - m_r^o)}{[1 + (\lambda_s - 1)m_r^o]} \quad (\text{A7})$$

$$z_{r,1} = f_{r,1} + f_{r,1} \frac{(\lambda_s - 1)(1 - m_r^o)}{[1 + (\lambda_s - 1)m_r^o]}. \quad (\text{A8})$$

Since we want the expected number of alleles shared *IBD*, we use $m_r = 2z_{r,2} + z_{r,1}$ to determine that

$$m_r = m_r^o + c_r \frac{(\lambda_s - 1)}{[1 + (\lambda_s - 1)m_r^o]} \quad (\text{A9})$$

where c_r is

$$c_r = 2f_{r,2}(2 - m_r^o) + f_{r,1}(1 - m_r^o). \quad (\text{A10})$$

Note that m_r depends on linkage only through the parameter λ_s ; the other terms, m_r^o and c_r , are computed under the null of no linkage, and hence depend only on the relationship of the pair of relatives.

Like others [4–7], we assume that λ_s is an exponential function of pair-specific covariates, so that $\lambda_s = \exp(X^T \beta)$, where X and β are possibly vectors for multiple covariates, and X includes an intercept, with corresponding coefficient β_0 . Substituting this exponential form into equation (A9) results in

$$m_r = m_r^o + c_r \frac{(\exp\{X^T \beta\} - 1)}{[1 + (\exp\{X^T \beta\} - 1)m_r^o]}. \quad (\text{A11})$$

To develop *QL* score statistics, we use a first-order Taylor series approximation of equation (A11). This is motivated by the facts that score statistics are based on the behavior of the score function near the null hypothesis and they are powerful for a sequence of alternative hypotheses that are contiguous to the null. Hence, using a first-order expansion around $\beta = 0$ results in

$$m_r = m_r^o + c_r X^T \beta. \quad (\text{A12})$$

Besides allowing us to view the expected allele sharing as a linear regression model with an offset of m_r^o and a scaled covariate, $X^* = c_r X$, this also simplifies the calculation of the *QL* score statistic in expression (1), because $\partial m_r / \partial \beta_j = c_r X_{rj} = X_{rj}^*$ (we drop the subscript for a pedigree, *r* is for the *r*-th relative pair in a pedigree and *j* is for the *j*-th covariate).

Alternative Covariate Scaling Factor

At this point, it is worthwhile to understand the impact of the assumption of no dominance variance. For *ASPs*, a multinomial logistic model has been used for the expected allele sharing proportions [4, 6], with the form

$$z_j(x_i) = \frac{\exp(X_i^T \beta_j)}{1 + \exp(X_i^T \beta_1) + \exp(X_i^T \beta_2)}. \quad (\text{A13})$$

Because constraints on the *IBD* sharing proportions can increase power, Greenwood and Bull used simulations to evaluate the impact of various constraints, versus an unconstrained mod-

el. They found that the assumption of no dominance variance, applied to all covariate configurations, can have greater power than the unconstrained model, and was a strong competitor to both the triangle constraint [36] and a minimax constraint [19]. For ASPs, the assumption of no dominance variance implies that $z_1 = 0.5$, which can be confirmed with Appendix equation (A8) by noting that $m_r^0 = 1$ for ASPs.

Because the multinomial logistic regression model in equation (A13) requires two sets of regression coefficients (β_1 and β_2), implying degrees of freedom that is twice the number of covariates, the power to detect linkage with this model can be weak. The assumption of no dominance variation reduces the number of parameters to a single set of regression coefficients. As an alternative to the assumption of no dominance variance, multiplicative models have been proposed for ASP data, whereby the covariate effect has twice the influence on the log relative risk for $IBD = 2$ versus $IBD = 1$ [8–10, 37]. Extending the multiplicative model to general types of relative pairs, the allele sharing probabilities can be expressed as

$$z_{r,2} = \frac{f_{r,2} \exp(2X^T \beta)}{f_{r,0} + f_{r,1} \exp(X^T \beta) + f_{r,2} \exp(2X^T \beta)}$$

$$z_{r,1} = \frac{f_{r,1} \exp(X^T \beta)}{f_{r,0} + f_{r,1} \exp(X^T \beta) + f_{r,2} \exp(2X^T \beta)}$$

Using $m_r = 2z_{r,2} + z_{r,1}$, and taking a first-order approximation about $\beta = 0$, it can be shown that the resulting scaling factor, c_r , for this multiplicative model is the same as that given in expression (A10) for the no dominance model.

Another model for IBD sharing probabilities was proposed by Goddard and Olson [1], who used the minimax constraint of Whittemore and Tu [19], which can be expressed as $z_1 = 0.335 + 0.58 z_0$. Applying this constraint to the multinomial regression for general types of affected relative pairs results in

$$z_{r,1} = \frac{f_{r,1} \exp(X^T \beta)}{f_{r,0} + f_{r,1} \exp(X^T \beta) + f_{r,2} [3.634 \exp(X^T \beta) - 2.643]}, \quad (A14)$$

$$z_{r,2} = \frac{f_{r,2} [3.634 \exp(X^T \beta) - 2.643]}{f_{r,0} + f_{r,1} \exp(X^T \beta) + f_{r,2} [3.634 \exp(X^T \beta) - 2.643]}. \quad (A15)$$

Then, using these sharing proportions in $m_r = 2z_{r,2} + z_{r,1}$, and taking a first-order approximation about $\beta = 0$, it can be shown that

$$m_r \approx m_r^0 + c_r X^T \beta, \quad (A16)$$

where the scaling factor now has the form

$$c_r = 7.268 f_{r,2} - 5.634 f_{r,2} f_{r,1} - 7.268 f_{r,2}^2 + f_{r,1} - f_{r,1}^2. \quad (A17)$$

Hence, by using c_r from equation (A17) to scale the covariates, we can approximate the minimax constraint used by Goddard and Olson [1] in our QL score statistics. Note that the scaling factor (A10), which results from assuming either no dominance variance or the multiplicative model, and the scaling factor (A17), which results from the minimax constraint, differ only for relative pairs that can share 2 alleles IBD , such as full sibs or double first-cousins; for other types of relative pairs, the two scaling factors are equivalent, and equal to $c_r = f_{r,1}(1 - f_{r,1})$. Hence, the impact of the choice of a particular scaling factor depends on the fraction of the sample that has affected relative pairs that can share 2 alleles IBD .

Appendix 2. Chi-bar-squared Mixture Proportions

Mixture proportions for null distributions of test statistics with multivariate inequality constraints are given in the Appendix of Wolak [38]. For our case of dimension three, these proportions are

$$w_3(\Sigma) = \frac{1}{4\pi} [2\pi - \arccos(\rho_{12}) - \arccos(\rho_{13}) - \arccos(\rho_{23})],$$

$$w_2(\Sigma) = \frac{1}{4\pi} [3\pi - \arccos(\rho_{12,3}) - \arccos(\rho_{13,2}) - \arccos(\rho_{23,1})],$$

$$w_1(\Sigma) = \frac{1}{2} - w_3(\Sigma),$$

$$w_0(\Sigma) = \frac{1}{2} - w_2(\Sigma),$$

where ρ_{ij} is the ij -th element of the correlation matrix associated with the covariance matrix Σ , and $\rho_{ij,k}$ is the partial correlation of X_i and X_j , with X_k fixed, assuming that the vector X has a multivariate normal distribution with covariance matrix Σ . As detailed elsewhere [39, pages 35–43],

$$\rho_{ij,k} = \frac{\rho_{ij} - \rho_{ik}\rho_{jk}}{\sqrt{1 - \rho_{ik}^2} \sqrt{1 - \rho_{jk}^2}}.$$

References

- Goddard KA, et al: Model-free linkage analysis with covariates confirms linkage of prostate cancer to chromosomes 1 and 4. *Am J Hum Genet* 2001;68:1197–1206.
- S.A.G.E., Statistical Analysis for Genetic Epidemiology. 2004, Statistical Solutions Ltd.: Cork, Ireland.
- Risch N: Linkage strategies for genetically complex traits. III The effect of marker polymorphism on analysis of affected relative pairs. *Am J Hum Genet* 1990;46:242–253.
- Olson JM: A general conditional-logistic model for affected-relative-pair linkage. *Am J Hum Genet* 1999;65:1760–1769.
- Greenwood CMT, Bull SB: Incorporation of covariates into genome scanning using sib-pair analysis in bipolar affective disorder. *Genet Epidemiol* 1997;14:635–640.
- Greenwood CMT, Bull SB: Analysis of affected sib pairs, with covariates – with and without constraints. *Am J Hum Genet* 1999; 64:871–885.
- Bull S, et al: Regression models for allele sharing: Analysis of accumulating data in affected sib pair studies. *Stat Med* 2002;21: 431–444.
- Rice J: Commentary: The role of meta-analysis in linkage studies of complex traits. *Am J Med Genet* 1997;74:112–114.
- Morton NE: Logarithm of odds (lods) for linkage in complex inheritance. *Proc Natl Acad Sci USA* 1996;93:3471–3476.

- 10 Rice J, et al: Covariates in linkage analysis. *Genet Epidemiol* 1999;17:691–695.
- 11 Zhang W, et al: A linkage tournament: Affection status, parametric analysis, multivariate traits, and enhancements to variance components and relative pairs. *Ann Hum Genet* 2002;66:87–98.
- 12 Schaid D, et al: Regression models for linkage: Issues of traits, covariates, heterogeneity, and interaction. *Hum Hered* 2003;55: 86–96.
- 13 Gauderman W, Siegmund K: Gene-environment interaction and affected sib pair linkage analysis. *Hum Hered* 2001;52:34–46.
- 14 Peng J, Tang H-K, Siegmund D: Genome scans with gene-covariate interaction. *Genet Epidemiol* 2005. To Appear.
- 15 McCullagh P, Nelder JA: Generalized linear models. London, Chapman and Hall, 1983.
- 16 Heyde C: Quasi-likelihood and its applications: A general approach to optimal parameter estimation. New York, Springer, 1997.
- 17 Bourgain C, et al: Novel case-control test in a founder population identifies P-selectin as an atopy-susceptibility locus. *Am J Hum Genet* 2003;73:612–626.
- 18 Risch N: Linkage strategies for genetically complex traits. II The power of affected relative pairs. *Am J Hum Genet* 1990;46:229–241.
- 19 Whittemore AS, Tu I-P: Simple, robust linkage tests for affected sib pairs. *Am J Hum Genet* 1998;62:1228–1242.
- 20 Dizier M, et al: The triangle test statistic (TTS): A test of genetic homogeneity using departure from the triangle constraints in IBD distribution among affected sib-pairs. *Ann Hum Genet* 2000;64:433–442.
- 21 Guo S-W: Gene-environment interactions and the affected-sib-pair designs. *Hum Hered* 2000;50:271–285.
- 22 Self SG, Liang K-Y: Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J Am Stat Assoc* 1987;82:605–610.
- 23 Blackwelder WC, Elston RC: A comparison of sib-pair linkage tests for disease susceptibility loci. *Genet Epidemiol* 1985;2:85–97.
- 24 Silvapulle M, Silvapulle P: A score test against one-sided alternatives. *J Am Stat Assoc* 1995; 90:342–349.
- 25 Wollan P, Dykstra R: Minimizing linear inequality constrained mahalanobis distances. *Appl Stat* 1987;36:234–240.
- 26 Abecasis G, et al: Merlin – rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 2002;30:97–101.
- 27 Cunningham JM, et al: Genome linkage screen for prostate cancer susceptibility loci: Results from the Mayo Clinic Familial Prostate Cancer Study. *Prostate* 2003;57:335–346.
- 28 Schaid D, et al: Comparison of microsatellites versus single nucleotide polymorphisms by a genome linkage screen for prostate cancer susceptibility loci. *Am J Hum Genet* 2004;75:948–965.
- 29 Kong A, Cox NJ: Allele-sharing models: LOD scores and accurate linkage tests. *Am J Hum Genet* 1997;61:1179–1188.
- 30 Guo X, et al: Small-sample performance of the robust score test and its modifications in generalized estimating equations. *Stat Med* 2005;24:3479–3495.
- 31 Greenwood C, Bull S: Down-weighting of multiple affected sib pairs leads to biased likelihood-ratio tests, under the assumption of no linkage. *Am J Hum Genet* 1999;64: 1248–1252.
- 32 Rotnitzky A, Jewell NP: Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* 1990;77:485–497.
- 33 Wallace C, et al: Linkage analysis using co-phenotypes in the BRIGHT study reveals novel potential susceptibility loci for hypertension. *Am J Hum Genet* 2006;79:323–331.
- 34 Lin DY: An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics* 2005;21:781–787.
- 35 Risch N: Linkage strategies for genetically complex traits. I Multilocus models. *Am J Hum Genet* 1990;46:222–228.
- 36 Holmans P: Asymptotic properties of affected-sib-pair linkage analysis. *Am J Hum Genet* 1993;52:362–374.
- 37 Holmans P: Detecting gene-gene interactions using affected sib pair analysis with covariates. *Hum Hered* 2002;53:92–102.
- 38 Wolak F: An exact test for multiple inequality and equality constraints in the linear regression model. *J Am Stat Assoc* 1987;82: 782–793.
- 39 Anderson T: An Introduction to Multivariate Analysis, ed 2. New York, John Wiley, 1984.