Published in final edited form as:

Biom J. 2010 February ; 52(1): 22–33. doi:10.1002/bimj.200900057.

Testing for Genetic Association in the Presence of Linkage and Gene Covariate Interactions

Andrea Callegaro^{*,1}, Jeremie J.P. Lebec¹, and Jeanine J. Houwing-Duistermaat¹

¹ Department of Medical Statistics and Bioinformatics, LUMC, 2300 RC Leiden, The Netherlands

Summary

In order to study family based association in the presence of linkage we extend a generalized linear mixed model proposed for genetic linkage analysis (Lebec and van Houwelingen, 2007) by adding a genotypic effect to the mean. The corresponding score test is a weighted FBAT statistic, where the weight depends on the linkage effect and on other genetic and shared environmental effects. For testing of genetic association in the presence of gene covariate interaction, we propose a linear regression method where the family-specific score statistic is regressed on family-specific covariates. Both statistics are straightforward to compute. Simulation results show that adjusting the weight for the within-family variance structure may be a powerful approach in the presence of environmental effects. The test statistic for genetic association in the presence of gene-covariate interaction improved the power for detecting association. For illustration we analyze the Rheumatoid Arthritis data from GAW15. Adjusting for smoking and anti-CCP increased the significance of the association with the DR locus.

Keywords

family-based studies; generalized linear mixed model; FBAT; Linkage; Linkage disequilibrium; Score test

1 Introduction

Hans van Houwelingen contributed to various types of analysis within the field of genetic epidemiology. In 1995, he adapted the goodness of fit test of le Cessie and van Houwelingen (1995) to a test for familial aggregation due to genetic effects (Houwing-Duistermaat et al., 1995). He has been interested in various measures for the analysis of data on family history (Houwing-Duistermaat and van Houwelingen, 1998; Jonker et al., 2003). For genetic linkage analysis he developed several score statistics (Putter et al., 2002; Lebec et al., 2004; Lebec and van Houwelingen, 2007). In this paper we will build further on his work on linkage analysis using generalized linear mixed models (Lebec and van Houwelingen, 2007). For genetic association analysis we will consider data on nuclear families. In the context of testing for association in the presence of linkage we will deal with two issues namely adjusting for the within family phenotypic variance and for the gene-covariate interaction. An example of a gene-covariate interaction is the interaction between HLA-DRB1 SE (shared epitope) gene and smoking in the development of rheumatoid arthritis

(Padyukov et al., 2004). The derived statistics appear to be extensions of the family based association tests (FBAT) statistics (Rabinowitz and Laird, 2000; Lake et al., 2000).

To identify new genetic factors in family data, dense single nucleotide polymorphisms (SNP) arrays are commonly used (Jaquish, 2007), necessitating the development of advanced methods for genetic association testing which model the complex correlation structure within families. One analysis strategy for family data is to first perform a linkage analysis followed by testing for genetic association in regions showing linkage. In this paper we develop methods for this second step, i.e. testing for genetic association in the presence of linkage. In several papers, it has been advocated to use a test for association which includes the information on identical by descent (IBD) sharing of the relatives. For quantitative traits, methods have been derived by Fulker et al. (1999) and Abecasis et al. (2000). For survival data a family based test was derived for the gamma frailty model by Zhong and Li (2004). Recently, for binary data a family based test for association using a gamma frailty and a log-log link was derived by Jonasdottir et al. (2007). In this paper we use the generalized linear mixed model framework to derive a test for association in the presence of linkage. These methods can be used for normally distributed, binary, count and survival data. Normally distributed random effects will be considered, because they allow for straightforward modelling of complex correlation structures.

Families are often ascertained based on phenotypic data. For example families should have at least two affected siblings. To deal with ascertainment, we will use the retrospective likelihood of the genotypes conditional on the phenotypes (Rabinowitz and Laird, 2000; Lake et al., 2000). We will assume that the marginal effects of covariates, the total variance and the heritability are known since it requires complex modelling to estimate these parameters in selected samples. Since in general a normally distributed random effect cannot be integrated out of the log likelihood function, we will derive the score statistic from an approximated likelihood approach (Breslow and Clayton, 1993). To obtain a test which is robust to population structure the conditional distribution of the siblings' genotypes given the parental genotypes will be used.

Identification of genetic factors which play a role in complex traits is hampered by the fact that phenotypes may be a result of different underlying biological mechanisms. Adjusting the phenotypes for covariates provides often only a small improvement of the efficiency of a test statistic (Whittemore et al., 2005). Modelling interaction between genetic factors and covariates may be more promising. Indeed for several complex disorders, genetic factors play only a role in a subset of the cases. One example is Rheumatoid Arthritis for which the underlying genetic mechanisms appear to be different in CCP positive compared to CCP negative patients (Van der Woude et al., 2009). If gene covariate interaction is present, the test statistic will have a different mean for different covariate values. Therefore we propose to regress the pseudo score statistic for genetic association on family specific covariates and to use a likelihood ratio statistic for testing.

The aim of this paper is to extend the FBAT statistic by considering the covariance structure of the sibling pairs (phenotypic variance) and by allowing for interaction effects between the SNP and a covariate. The performance of the test statistics will be studied via simulations. Secondly, we apply the methods to the simulated GAW15 samples (Witte et al., 2007) to test for association and we will compare the results with other tests which have been applied to these simulated data. Finally, the test statistics will be applied to real data from the NARAC study.

2 The Generalized Linear Mixed Model

We use the same formulation as in Lebrek and van Houwelingen (2007). Let $y = (y_1, \dots, y_m)$ be the trait values of m siblings. Let Z be the $m \times k$ matrix containing the values for k covariates of the m siblings. Let $b = (b_1, \dots, b_m)$ be a vector of random effects with zero mean and variance σ^2 . These random effects represent genetic and shared environmental effects and conditional on these random effects and the vector of covariate effects β , the outcomes of the subjects are assumed to be independently distributed. The density function f of the exponential family will be used and the log likelihood conditional on the random effects b_j can be written as follows

$$\log f(y_j | \beta, b_j) = y_j \times (z_j^T \beta + b_j) + a(y_j) - \psi(z_j^T \beta + b_j),$$

where the first and the second derivatives of ψ determine the first and the second moment of y_j , i.e. $\psi'(z_j^T \beta + b_j) = E(y_j | \beta, b_j)$ and $\psi''(z_j^T \beta + b_j) = \text{var}(y_j | \beta, b_j)$. For normally distributed data the identity link is used. In the case of binary traits, the logit link is used i.e. $\psi'(x) = \exp(x) / (1 + \exp(x))$ and $\psi''(x) = \psi'(x)(1 - \psi'(x))$. For count data the Poisson model should be used i.e. $\psi'(x) = \psi''(x) = \exp(x)$. For survival data (age at onset) the Poisson model can be used by substituting the term $\exp(x)$ with the baseline cumulative hazard $(\Lambda_0(t) \exp(z^T \beta))$. Note that the baseline cumulative hazard Λ_0 and the parameters of a logistic or Poisson model depend on the variance of the random effect.

To obtain the log likelihood function, multivariate integrals over the distribution of b have to be considered. Since for binary and survival outcomes no closed formula exist, we will use a quasi likelihood approach. Using a second order Taylor approximation of the log likelihood around $b = 0$ Lebrek and van Houwelingen (2007) showed that the log likelihood function behaves like the normal log likelihood based on the following normal distribution

$$y \sim N(\psi'(Z\beta), \Sigma_y(\gamma)),$$

where $\psi'(Z\beta)$ denotes the vector whose j th element is given by $\psi'(z_j^T \beta)$ and $\Sigma_y(\gamma) = \psi''(Z\beta) + \psi''(Z\beta)\Sigma(\gamma)\psi''(Z\beta)$, with $\psi''(Z\beta)$ the diagonal matrix whose j th diagonal element is given by the scalar $\psi''(z_j^T \beta)$.

If linkage at a locus is present, the covariance between two siblings depends on their IBD status at the locus. The following covariance structure between the random effects of two relatives j and k is used (Lebrek and van Houwelingen, 2007)

$$[\Sigma(\gamma)]_{jk} = \sigma^2 \begin{cases} 1, & \text{if } j=k; \\ (\pi_{jk} - E\pi_{jk})\gamma + \rho, & \text{if } j \neq k, \end{cases}$$

where π is the proportion of alleles shared identical by descent (IBD) between siblings at the trait locus; γ is the linkage effect and ρ is the intra class correlation between the relatives at the population level. In this paper we assume that the marginal population parameters (ρ, σ^2) are known from previous twin/family studies.

For genetic association testing, the SNP genotypes have to be included in the model. Let $g = (g_1, \dots, g_m)$ be the genotypes of the m siblings and $g_p = (g_m, g_l)$ be the genotypes of the

parents. Only diallelic SNPs will be considered. Let $X_g = 0, 1$ or 2 be the number of the SNP's less frequent (minor) allele. To model association we propose the following distribution

$$\mathbf{y} \sim N\left(\psi'(\mathbf{Z}\beta + X_g\beta_g), \Sigma_y(\gamma)\right), \quad (1)$$

where β_g represents the effect of the SNP on the phenotype. Note that by including the term X_g only in the argument of ψ' (but not in the argument of ψ'') the model is not exponential anymore. We use this simplified model in order to derive simple weights for FBAT. Further, we will show in section 4.1 that the weights derived from model in equation (1) are reasonable and they are similar to the weights derived by other authors (Houwing-Duistermaat et al., 2007; Jonasdottir et al., 2007; Zhong and Li, 2004).

Now the retrospective likelihood for the sibship can be written as

$$P(\mathbf{g}|\mathbf{y}, \mathbf{g}_p) = \frac{P(\mathbf{g}|\mathbf{g}_p)P(\mathbf{y}|\mathbf{g})}{\sum_g P(\mathbf{g}|\mathbf{g}_p)P(\mathbf{y}|\mathbf{g})}$$

where $\sum_g$ denotes summation over all possible vectors of genotypes $\mathbf{g}$ for the sibship.

The retrospective quasi score function for the i th sibship $i = 1, \dots, N$ is given by

$$\begin{aligned} U_i(\gamma) &= \frac{\partial \ell(\beta_g, \gamma)}{\partial \beta_g}(\beta_g=0) \\ &= \left[X_{g,i} - E(X_{g,i}|\mathbf{g}_{p,i}) \right]^T \Psi''(\mathbf{Z}_i\beta) \Sigma_{y,i}^{-1}(\gamma) (\mathbf{y}_i - \psi'(\mathbf{Z}_i\beta)) \\ &= \left[X_{g,i} - E(X_{g,i}|\mathbf{g}_{p,i}) \right]^T \mathbf{W}_i(\gamma) (\mathbf{y}_i - \psi'(\mathbf{Z}_i\beta)) \end{aligned} \quad (2)$$

where $E(X_{g,i}|\mathbf{g}_{p,i})$ is the expected value of $X_{g,i}$ given the SNP genotypes of the parents. The linkage parameter (γ) is usually unknown and a nuisance parameter for the score test. Since $\ell(\beta_g=0, \gamma)$ is equal to zero, it is not needed to adjust the score function for the estimation of the nuisance parameter γ (Zhong and Li, 2004). Note that the score U is a weighted FBAT statistic (Rabinowitz and Laird, 2000; Lake et al., 2000) where the weights W depend on the environmental/genetic correlation and on the IBD status of the siblings.

When the IBD status for all sibling pairs is known, the variance of U can be computed:

$$\text{Var}(U_i(\gamma)) = (\mathbf{y}_i - \psi'(\mathbf{Z}_i\beta))^T \mathbf{W}_i(\gamma) \times \text{Var}_{\text{IBD}}(X_{g,i} - E(X_{g,i}|\mathbf{g}_{p,i})) \times \mathbf{W}_i(\gamma) (\mathbf{y}_i - \psi'(\mathbf{Z}_i\beta)).$$

The variances of $[X_{g,i} - E(X_{g,i}|\mathbf{g}_{p,i})]$ are given in table 1 as function of the parental genotypes and the IBD status of the sibling pairs. The test statistic is given by

$$S_{\text{IBD}} = \frac{\sum_{i=1}^N U_i(\gamma)}{\sqrt{\sum_{i=1}^N \text{Var}(U_i(\gamma))}} \quad (3)$$

When the IBD status is unknown S_{IBD} cannot be computed. In this case the variance can be empirically estimated and the test statistic for association in the presence of linkage is given by

$$S_e = \frac{\sum_{i=1}^N U_i(\gamma)}{\sqrt{\sum_{i=1}^N U_i(\gamma)^2}} \quad (4)$$

Both statistics (3) and (4) are normally distributed with mean zero and variance one under the null hypothesis of no association. Lake et al. (2000) showed that the tests with empirical variance-covariance estimator are valid for association in presence of linkage. Another important property of the derived statistic is that the mean is null under the null hypothesis, irrespectively of the weight functions, i.e. for the model used for y . This means that the test is valid under the null hypothesis even though the weights may not be optimal. Finally, note that the empirical variance FBAT (FBATe) proposed by Lake et al. (2000) corresponds to the statistic in equation (4) where $\mathbf{W}(\gamma)$ in the score (2) is the identity matrix.

3 Adjusting for Gene-Covariate Interaction

The second aim of this paper is to deal with gene covariate interaction. Let $Z_{F_c,i}$ be a standardized family-specific covariate with mean zero and variance one. One example of a family-specific covariate is the sum of sibling specific covariates. The following FBAT linear regression is proposed

$$U_i(\gamma) = \mu_g + \sum_{c=1}^C \mu_c Z_{F_c,i} + \epsilon_i, \quad (5)$$

where $U_i(\gamma)$ is the contribution of the i th family to the weighted FBAT score statistic given by equation (2) and $\epsilon_i \sim N(0, \sigma_\epsilon^2)$. To test the null hypothesis of no association, we propose to use the likelihood ratio statistic for $\mu_g = 0$ (LR_{adj}). The coefficients of the covariates (μ_c) are considered as nuisance parameters. The statistic tests for a difference from zero in families with "average" covariates. This approach is similar in spirit to the method of Kong and Cox (1997) for linkage analysis, which is a "regression" of the score statistics. The likelihood ratio test for $\mu_g = 0$ when no family-specific covariates are considered ($\mu_c = 0$, $c = 1, \dots, C$) is a test for association given linkage ignoring the gene-covariate interactions (LR).

4 Results

4.1 Weight functions

To demonstrate how the weight function in equation (2) depends on the IBD sharing we present here examples for binary and age at onset (survival) traits.

For binary traits the weights as function of the number of alleles shared IBD are given in figure 1 for affected sibling pairs (ASP) as well as for discordant sibling pairs (DSP). A sib-sib marginal correlation ρ of 0.5, variance σ^2 of 2 and a linkage effect γ of 1 was used in the weight. Two values for the prevalence of the trait were used namely 0.1 and 0.2. The parameter β_0 was obtained from the marginal prevalence and the variance σ^2 . No covariates

were included. The weights become more extreme with higher prevalences of the disease and for affected sibling-pairs they decrease with the number of alleles shared IBD.

For survival data, we used the same setting as described by Zhong and Li (2004). Briefly, a Weibull baseline hazard function $\Lambda_0(t) = (t/d)^\tau$, with $d = 80$ and $\tau = 5$ was used and a gamma distributed random effect with mean 1, variance equal to $\sigma_c^2 = 3.3$ and marginal sib-sib correlation equal to $\rho = 0.5$.

In the weights a variance of the normal distributed random effect of $\sigma_N^2 = \log(1+3.3)$ was used and the marginal hazard was used as baseline hazard. Figure 2(a) shows the corresponding population disease-free survival curve. Figure 2(b)-2(d) show for affected sibling pairs the weights as function of their age at onset for the three IBD categories. As expected, a large weight is assigned to early-onset siblings and weights decrease with the number of alleles shared IBD.

4.2 Simulation study

To compare the performance of the FBAT_e statistic and the new statistics S_{IBD} , S_e , LR and LR_{adj} we performed a simulation study. Data-sets of 300 nuclear families with two siblings were simulated. Two designs were considered. One with 300 discordant sibling pairs (DSP) and one with 150 affected sibling pairs (ASP) and 150 DSPs. The data for each family were simulated using the following five steps: 1) Generate parental disease locus genotypes assuming random mating and Hardy-Weinberg genotype frequencies, with variant allele frequency equal to 10%. 2) Given the parental genotype, generate offspring genotypes assuming Mendelian transmission. 3) Generate offsprings exposure indicators (Z) with an exposure prevalence of 20% and normally distributed random effects (b), regardless of the genotype. Let r denote the portion of the random effects variance explained by the shared random effects. 4) Given the offspring genotypes, covariates and random effects, generate offspring disease phenotypes according to the following logistic model

$$P(y=1|X_g, Z, b) = \frac{\exp(\beta_0 + \beta_g X_g + \beta_z Z + \beta_{gz} X_g Z + b)}{1 + \exp(\beta_0 + \beta_g X_g + \beta_z Z + \beta_{gz} X_g Z + b)} \quad (6)$$

where X_g is the sum of the variant alleles. 5) Generate the parental biallelic marker data using a logistic model, with X_g as covariate. We assumed that no recombination occurs between the locus and the marker and that the marker and the disease locus were in high linkage disequilibrium.

We estimated the type one error rate based on 5000 simulations under independence between marker and disease locus. For all models and tests considered the actual significance level agreed with the nominal level (data not shown).

Table 2 shows the results of data simulated with and without the presence of shared environmental effects. Phenotypic data were generated using the logistic model (6) with $\beta_z = \beta_{gz} = 0$ and with independent random effects ($r = 0$) or with shared random effects ($r = 0.4$). The prevalence of the disease was 20% with linkage effect $\gamma = 0.025$. Table 2 shows that S_e has slightly more power with respect to FBAT_e in the presence of shared environmental effects and similar power when no shared environmental effects are present. Further, these results show that positive environmental correlation decreases the required sample sizes for DSP (Lange et al., 2002) and increased the required sample sizes for the pooled set of ASPs and DSPs.

Table 3 shows the results of the simulations with and without gene-covariate interaction. The following parameters in model (6) were used $\beta_0 = -2.75$, $\beta_g = 0.76$, $\beta_z = 1$, $\beta_{gz} = (0, 2, -2)$ and $\text{var}(b) = 0$. The prevalence of the disease was 6% with 10% for the exposed ($Z = 1$) and 5% for the unexposed ($Z = 0$). For these simulations the test statistic FBAT_e and S_e performed similar because no shared environmental effects were present. The test statistic S_{IBD} performed slightly better for the pooled set of ASPs and DSPs. The likelihood ratio tests were always more powerful than the score test FBAT_e . Finally, adjusting for the interaction between genotype and covariate improved the power if interaction is present. In the case of no interaction the LR_{adj} approach had similar power compared to the unadjusted version (LR).

4.3 Simulated GAW15 data-sets

We applied the proposed approaches to the 100 GAW15 simulated data sets. The simulation model is described in Witte et al. (2007). Briefly, a lifetime prevalence of 0.01 and a lambda-sib (lifetime prevalence for siblings of affected individuals divided by the population prevalence) of 9.03 was used. For the DR locus two risk variants (DR1 and DR4) were simulated. The DR1 allele had a frequency of 0.1 and a genotype relative risk for homozygous carriers versus homozygous carriers of the DRx allele of 1.5. The other variant DR4 had a frequency of 0.25 and increased the risk for RA enormously. The genotype relative risk for homozygous carriers DR4 versus homozygous carriers of the DRx allele was 30. We studied the performance of the new statistics to identify the DR1 allele using the same selection criteria as in Houwing-Duistermaat et al. (2007), i.e. sibling pairs who were homozygous carriers of the DR4 allele were removed. The number of affected sibling pairs used for analysis varied from 211 to 270 with a mean number of 238 sibling pairs. The average number of informative families for family based association was 100 per data-set.

The expected values of the number of minor alleles conditioned on the parental genotypes were computed using the FBAT package (Lake et al., 2000). The prevalence of RA (0.01) and the overall sibling relative risk (9.03) corresponds to a covariance $\rho\sigma_N^2 = 0.75$. Based on an arbitrary value $\rho = 0.5$ the corresponding variance of the random effect is $\sigma_N^2 = 1.5$ (see Lebec and van Houwelingen (2007) for more details). The mean value -over the 100 replications- of the linkage parameters estimated by the Lebec and van Houwelingen (2007) model with these parameters was $\gamma = 0.2$.

The results of the association analysis of the DR1 allele are given in table 4. At the DR locus, the proposed weighted score test (S_e) is equivalent to the standard test statistic FBAT_e . We also applied the regression approach FBAT_{adj} . We adjusted simultaneously for the covariates sex, smoke, anti-CCP, Igm, severity and age at onset. Adjusting for the covariates increased the number of rejections of the null hypothesis from 45 times to 61 times.

4.4 Application to the NARAC data

We applied the proposed methods to the North American Rheumatoid Arthritis Consortium study (NARAC). Of the four GAW15 studies on Rheumatoid Arthritis (RA), only the NARAC data set was suitable for this analysis because at least some parental HLA-DRB1 genotypes were available. Following the work of Newton et al. (2004) we coded HLA-DRB1 into two distinct alleles. The shared epitope (SE) alleles DRB1*0401, *0404, *0405, *0408, *0101, *0102, *1402, *09 and *1001 and the not SE alleles (N). It is known that 30% of the patients do not carry SE alleles. HLA-DRB1 error rate was less than 1%. Individuals' genotypes that were in error were omitted from the analysis. Most of the families were nuclear families, however there were a few extended families. These large

families were splitted into nuclear families. 735 nuclear families with at least two affected siblings were selected for analysis. Of these families, 615 had two affected siblings, 91 had 3 affected siblings, 23 had 4 affected siblings and 4 and 2 families had 5 and 6 affected siblings, respectively.

First a Kong and Cox linkage analysis (Kong and Cox, 1997) was performed. A LOD score of 20 was obtained at the estimated using Kong and Cox in the correspondence of the HLA-DRB1 gene locus. For the computation of the weights, a prevalence of 0.01, a sib-sib marginal correlation of $\rho = 0.5$ and a variance of $\sigma_N^2 = 5$ were used (Lebrech et al., 2007). The linkage parameter estimated by the Lebrech and van Houwelingen (2007) model with these parameters was $\gamma = 0.3$. For family-based association analysis, only 51 families were informative. Results of the association analysis are presented in table 5. The standard FBAT_e gave a p-value of 8.2e-06. The new S_e statistic gave higher p-values (p-value= 6.5e-05). The corresponding likelihood ratio statistic yielded a p-value of 5.2e-07. For testing in the presence of gene covariate interaction, smoking status and anti-CCP were considered. As illustration we also adjusted for all available covariates namely sex, age at onset, anti-CCP, smoke, the number of subcutaneous nodules, RF, the number of tender joints, the number of swollen joints, and the left hand erosions. Because of missing covariate values, 43 families were available. Table 5 shows the results for the most important covariates. Figure 3 shows the family-specific contribution to the FBAT score test versus the family specific covariates. Adjusting for the covariates decreased the p-value. This analysis suggests that anti-CCP and smoke are the most important covariates.

5 Discussion

We proposed new test statistics for family based association testing in the presence of linkage and gene covariate interaction for general phenotypes. Using the generalized linear mixed model by Lebrech and van Houwelingen (2007), we derived a weighted FBAT statistic where the weight depends on the covariances between siblings due to polygenic and shared environmental effects, and on the number of alleles shared IBD (linkage effect). This new statistic generalized the statistic of Lunetta et al. (2000) by including the phenotypic variance-covariance structure. For gene-covariate interaction, we proposed a simple regression model of the FBAT statistic on the covariate. Here a likelihood ratio statistic was used.

For no shared environmental effects, the weights of our test statistic for survival and binary data are similar to the weights of the model of Zhong and Li (2004) and Jonasdottir et al. (2007), respectively. Our simulation results showed that, for common disease, the weighted FBAT is more powerful than the standard FBAT especially in the presence of environmental effects and for DSP. Adjusting for linkage effect when shared environmental effects were not present (on the variance-covariance structure) only slight increased the power. In fact, for prevalences smaller than 0.1 the weights are almost independent of the IBD status. These results agree with the simulation results of Zhong and Li (2004). Jonasdottir et al. (2007) showed that adjusting for linkage effect improved substantially the power. However, they compared their method with the "optimal" offset FBAT (Lunetta et al., 2000) while we compared our method with the standard FBAT. Lange et al. (2002) showed that the "optimal" offset FBAT may not be powerful for selected samples.

In this paper we assumed that the marginal population parameters (ρ, σ^2) are known. In principle, one needs random sets of families from the population to estimate those parameters. A general method on how to obtain sensible estimates is described in Lebrech and van Houwelingen (2007). For binary traits, one simply needs data on prevalence and recurrence risk ratio. When these parameters are known, our simulation results showed that

S_e is more powerful than $FBAT_e$ in the presence of shared environmental effects. Computing the variance of the score S_{IBD} when identical by descent status is known further improved the power. On the GAW15 simulated data, the weighted and the unweighted FBAT performed similarly, probably due to the selection scheme.

Several population based methods for testing association in the presence of linkage are available, such as the score statistic proposed by Houwing-Duistermaat et al. (2007) and the LAMP (Li et al., 2005) method. These two methods were applied to the same set of 100 simulated GAW 15 data-set. Probably because these population based methods used also the information on available controls, the power of these methods was higher than the statistics considered in this paper. The main advantage of the family-based approaches is that they are robust against spurious association, which can arise with population substructure. Including independent controls in the GLM framework will be topic of future research.

For association testing in the presence of gene-covariate interaction, we proposed a simple linear regression method where the family-based contribution to the weighted (or unweighted) FBAT score statistic is regressed on family-based covariates. A standard procedure to test for association adjusting for C interactions is to use a $C + 1$ degrees of freedom test (Lunetta et al., 2000). When considering multiple covariates or when the interaction effect is small this test-statistic will lose power due to the increase in the number of degrees of freedom. In this paper we proposed an alternative one degree of freedom test. Our simulation results show that when gene-covariate interactions are present our method is more powerful than the standard FBAT statistics. When no gene-covariate interactions are present the power is not affected by adjusting for covariates. On the GAW15 data (real and simulated), our approach outperforms the standard $FBAT_e$. LR_{adj} is a very promising method because in these simulated data-sets, Chen et al. (2007) showed that $FBAT_e$ is more powerful than other family-based methods such as conditional logistic regression method (Millstein et al., 2005), the parametric likelihood-based approach (LAMP) of Li et al. (2005) and the homozygote sharing test (Dupuis and Van Eerdewegh, 2003). Also for the NARAC data adjusting the score statistic for smoking and anti-CCP improved the significance of the test-statistic for genetic association at the DR locus. Recently Klareskog et al. (2006) showed that the risk from smoking for RA is particularly high among individuals who have a shared epitope allele and who also have elevated levels of anti-CCP, which is confirmed by our analysis.

To conclude, we extended the $FBAT_e$ test to adjust for the within-family variance structure and to allow for gene-covariates interaction. Especially the second adjustment improved power and we showed by simulations that no power is lost in the case of no gene-covariate interaction.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

AC, JJPL and JJHD were supported by a grant from the Netherlands Organization for Scientific Research (NWO 917.66.344). Support for generation of the simulated data was provided from NIH grants 5R01-HL049609-14, 1R01-AG021917-01A1, the University of Minnesota, and the Minnesota Supercomputing Institute. We would like to acknowledge also the GAW grant, R01-GM031575, as well as AR44422.

References

- Abecasis GR, Cardon LR, Cookson WO. A general test of association for quantitative traits in nuclear families. *American Journal of Human Genetics*. 2000; 66:279–292. [PubMed: 10631157]
- Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*. 1993; 88:9–25.
- Chen MH, Cui J, Guo CY, Cupples LA, Van Eerdewegh P, Dupuis J, Yang Q. Joint modeling of linkage and association using affected sib-pair data. *BMC proceedings*. 2007; 1:S38. [PubMed: 18466536]
- Dupuis J, Van Eerdewegh P. Identification of polymorphisms that explain a linkage peak: conditioning on parental genotypes [abstract]. *Genetic Epidemiology*. 2003; 25:247.
- Fulker DW, Cherny SS, Sham PC, Hewitt JK. Combined linkage and association sib-pair analysis for quantitative traits. *American Journal of Human Genetics*. 1999; 64:259–267. [PubMed: 9915965]
- Houwing-Duistermaat JJ, Derkx BHF, Rosendaal FR, van Houwelingen JC. Testing familial aggregation. *Biometrics*. 1995; 51:1292–1301. [PubMed: 8589223]
- Houwing-Duistermaat JJ, Uh HW, van Houwelingen HC. A new score statistic to test for association given linkage in affected sibling pair-control designs. *BMC Proceedings*. 2007; 1:S39. [PubMed: 18466537]
- Houwing-Duistermaat JJ, van Houwelingen HC. Incorporation of family history in logistic regression models. *Statistics in Medicine*. 1998; 17:2865–2882. [PubMed: 9921607]
- Jaquish CE. The Framingham Heart Study, on its way to becoming the gold standard for cardiovascular genetic epidemiology? *BMC Medical Genetics*. 2007; 8:63. [PubMed: 17916250]
- Jonasdottir G, Humphreys K, Palmgren J. Testing association in the presence of linkage - a powerful score for binary traits. *Genetic Epidemiology*. 2007; 31:528–540. [PubMed: 17436308]
- Jonker MA, Jacobi CE, Hoogendoorn WE, Nagelkerke NJD, de Bock GH, van Houwelingen JC. Modeling familial clustered breast cancer using published data. *American Association for Cancer Research*. 2003; 12:1479–1485.
- Klareskog L, Stolt P, Lundberg K, Killberg H, Bengtsson C, Grunewald J, Rnnelid J, Harris HE, K. UA, Rantap-Dahlqvist S, Eklund A, Padyukov L, Alfredsson L. A new model for an etiology of rheumatoid arthritis. smoking may trigger HLA-DR (shared epitope)-restricted immune reactions to autoantigens modified by citrullination. *Arthritis and Rheumatism*. 2006; 54:38–46. [PubMed: 16385494]
- Kong A, Cox NJ. Allele-sharing models: Lod scores and accurate linkage tests. *American Journal of Human Genetics*. 1997; 61:1179–1188. [PubMed: 9345087]
- Lake SL, Blacker D, Laird NM. Family-based tests of association in the presence of linkage. *American Journal of Human Genetics*. 2000; 67:1515–1525. [PubMed: 11058432]
- Lange C, DeMeo DL, Laird NM. Power and design considerations for a general class of family-based association tests: quantitative traits. *American Journal of Human Genetics*. 2002; 71:1330–1341. [PubMed: 12454799]
- le Cessie S, van Houwelingen JC. Testing the fit of a regression-model via score tests in random effects models. *Biometrics*. 1995; 51:600–614. [PubMed: 7662848]
- Lebre J, Putter H, van Houwelingen JC. Score test for detecting linkage to complex traits in selected samples. *Genetic Epidemiology*. 2004; 27:97–108. [PubMed: 15305326]
- Lebre JJ, Helmer Q, Nishchenko I, van Houwelingen HC. Adjusting for sex and anti-ccp levels in linkage analysis of rheumatoid arthritis. *BMC Proceedings*. 2007; 1:S75. [PubMed: 18466577]
- Lebre JJP, van Houwelingen HC. Score test for linkage in generalized linear models. *Human Heredity*. 2007; 64:5–15. [PubMed: 17483592]
- Li M, Boehnke M, Abecasis GR. Joint modeling of linkage and association: identifying SNPs responsible for a linkage signal. *American Journal of Human Genetics*. 2005; 76:934–949. [PubMed: 15877278]
- Lunetta KL, Farone SV, Biederman J, Laird NM. Family based tests of association and linkage using unaffected sibs, covariates and interactions. *American Journal of Human Genetics*. 2000; 66:605–614. [PubMed: 10677320]

- Millstein J, Siegmund KD, Conti DV, Gauderman WJ. Testing association and linkage using affected-sib-parent study designs. *Genetic Epidemiology*. 2005; 29:225–233. [PubMed: 16121357]
- Newton JL, Harney SM, Wordsworth BP, Brown MA. A review of the MHC genetics of rheumatoid arthritis. *Genes and Immunity*. 2004; 5:151–157. [PubMed: 14749714]
- Padyukov L, Silva C, Stolt P, Alfredsson L, Klareskog L. A gene-environment interaction between smoking and shared epitope genes in HLA-DR provides a high risk of seropositive rheumatoid arthritis. *Arthritis Rheumatism*. 2004; 50:3085–3092. [PubMed: 15476204]
- Putter H, Sandkuijl LA, van Houwelingen JC. Score test for detecting linkage to quantitative traits. *Genetic Epidemiology*. 2002; 22:345–55. [PubMed: 11984866]
- Rabinowitz D, Laird N. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Human Heredity*. 2000; 50:211–223. [PubMed: 10782012]
- Van der Woude D, Houwing-Duistermaat JJ, Toes REM, Huizinga TWJ, Thomson W, Worthington J, Van der Helm-van Mil AHM, de Vries RRP. Quantitative heritability of anti-citrullinated protein antibody-positive and anti-citrullinated protein antibody-negative rheumatoid arthritis. *Arthritis Rheumatism*. 2009; 60:916–923. [PubMed: 19333951]
- Whittemore AS, Halpern J, Ahsan H. Covariate adjustment in family-based association studies. *Genetic Epidemiology*. 2005; 28:244–255. [PubMed: 15593089]
- Witte JS, Schnell AH, Cordell HJ, Spielman RS, Amos CI, Miller MB, Almasy L, MacCluer JW. Introduction to Genetic Analysis Workshop 15 summaries. *Genetic Epidemiology*. 2007; 31:S1–6. [PubMed: 18046756]
- Zhong X, Li H. Score tests of genetic association in the presence of linkage based on the additive genetic gamma frailty model. *Biostatistics*. 2004; 5:307–327. [PubMed: 15054033]

Fig. 1. Number of alleles shared IBD versus weights for affected siblings with an affected sibling (a) and with a healthy sibling (b). The covariance parameters are $\sigma^2=2$, $\rho=0.5$ and $\gamma=1$. Straight and dashed lines represent the weights for diseases with prevalence of 20% and 10%, respectively.

Fig. 2. Disease free survival (a) and weights versus age at onset of sibling 1 for IBD=0 (b), IBD=1 (c) and IBD=2 (d). Covariance parameters are $\sigma^2 = 1.45$; $\rho = 0.5$ and $\gamma = 1$. Different lines represent different age at onset of the second sibling.

Fig. 3. NARAC data: Family-specific contribution of the unweighted FBAT test (y-axis) versus family-specific covariates. Dashed and dot lines represent the predicted values of the linear regression with and without covariate, respectively.

Table 1

Covariances between the genotypes of two siblings given their IBD status and the parental genotype.

Parental genotype	IBD		
	0	1	2
just one heterozygous parent	-0.25	0	0.25
both parents heterozygous	-0.5	0	0.5

Table 2

Power comparisons based on data simulated with and without environmental effect. Power estimated by 1000 simulations of 300 nuclear families. r is the portion of the variance explained by the shared environmental effect. p is the allele frequency of the disease allele. Significant level: $\alpha = 0.001$.

Method	DSP		DSP+ASP	
	$r = 0$	$r = .4$	$r = 0$	$r = .4$
$p = 0.1$:				
FBATe	0.36	0.57	0.71	0.58
S_e	0.37	0.60	0.71	0.61
S_{IBD}	0.37	0.60	0.72	0.62
$p = 0.3$:				
FBATe	0.31	0.44	0.54	0.44
S_e	0.31	0.46	0.54	0.47
S_{IBD}	0.32	0.46	0.55	0.48

Table 3

Power comparisons based on data simulated with and without gene-covariate interaction. Power estimated by 1000 simulations of 300 nuclear families. p is the allele frequency of the disease allele. Results of S_e are the same as FBAT_e. Significant level: $\alpha = 0.001$;

Method	DSP		DSP+ASP	
	$\beta_{g_{\pm}} = 0$	$\beta_{g_{\pm}} = 2$	$\beta_{g_{\pm}} = 0$	$\beta_{g_{\pm}} = 2$
$p = 0.1$:				
FBATe	0.36	0.88	0.24	0.93
S _{IBD}	0.36	0.86	0.25	0.94
LR	0.37	0.89	0.26	0.94
LR _{adj}	0.38	0.90	0.29	0.96

Table 4

GAW15 simulated data: number of times the null hypothesis is rejected using various test statistics for genetic association.

replicates	significance level	Method		
		FBAT_e	LR	LR_{adj}
1 to 100	0.05	95	95	95
1 to 100	0.001	45	52	61

Table 5

Results of analysis of NARAC data: P-values corresponding to likelihood ratio test statistic for genetic association of the HLA-DRB1 allele in the presence of linkage and gene-covariate interaction.

model	P-values
μ_g	1.8e-06
μ_g + anti-CCP	8.8e-07
μ_g + smoke	9.4e-07
μ_g + anti-CCP+smoke	4.0e-07