Genetic Epidemiology 30: 570–581 (2006)

# Semiparametric Variance-Component Models for Linkage and Association Analyses of Censored Trait Data

**G. Diao and D.Y. Lin***

*Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina*

Variance-component (VC) models are widely used for linkage and association mapping of quantitative trait loci in general human pedigrees. Traditional VC methods assume that the trait values within a family follow a multivariate normal distribution and are fully observed. These assumptions are violated if the trait data contain censored observations. When the trait pertains to age at onset of disease, censoring is inevitable because of loss to follow-up and limited study duration. Censoring also arises when the trait assay cannot detect values below (or above) certain thresholds. The latent trait values tend to have a complex distribution. Applying traditional VC methods to censored trait data would inflate type I error and reduce power. We present valid and powerful methods for the linkage and association analyses of censored trait data. Our methods are based on a novel class of semiparametric VC models, which allows an arbitrary distribution for the latent trait values. We construct appropriate likelihood for the observed data, which may contain left or right censored observations. The maximum likelihood estimators are approximately unbiased, normally distributed, and statistically efficient. We develop stable and efficient numerical algorithms to implement the corresponding inference procedures. Extensive simulation studies demonstrate that the proposed methods outperform the existing ones in practical situations. We provide an application to the age at onset of alcohol dependence data from the Collaborative Study on the Genetics of Alcoholism. A computer program is freely available. *Genet. Epidemiol.* 30:570–581, 2006. © 2006 Wiley-Liss, Inc.

**Key words: age at onset; censoring; complex diseases; IBD sharing; linkage disequilibrium; LOD score; maximum likelihood; proportional hazards; quantitative traits; random effects; transformation**

Contract grant sponsor: National Institutes of Health (NIH).
*Correspondence to: Danyu Lin, Ph.D., Department of Biostatistics, University of North Carolina, McGavran-Greenberg Hall, CB #7420, Chapel Hill, NC 27599-7420. E-mail: lin@bios.unc.edu
Received 16 February 2006; Revised 12 April 2006; Accepted 15 May 2006
Published online 20 July 2006 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/gepi.20168

## INTRODUCTION

The study of complex diseases is the most important challenge in genetic epidemiology. Because complex diseases are characterized by various quantitative traits, QTL mapping plays a critical role in this endeavor. The most popular approach to QTL mapping pertains to variance-component (VC) models [Amos, 1994; Amos et al., 1996; Almasy and Blangero, 1998; Fulker et al., 1999; Abecasis et al., 2000]. This approach is attractive because it accommodates any type of pedigree, allows both linkage and association analyses, and tends to be more powerful than competing methods.

In many studies, the trait data contain censored observations. Trait censoring can arise in several ways. For example, the trait assay may fail to detect values below (or above) certain thresholds. This is the case with the coronary artery calcification (CAC) data in the Family Heart Study [Higgins et al., 1996]: the distribution of CAC exhibits a spike at the left end because a large proportion of CAC measures do not exceed threshold for detection and are thus recorded as 0; the non-zero CAC measures are skewed to the right. A similar study was described by Epstein et al. [2003]. Trait censoring may also arise from subject-specific thresholds due to factors such as medication [Valle et al., 1998; Epstein et al., 2003].

Another type of censoring occurs when the quantitative trait pertains to event time, such as age at onset of disease or survival time. Event times are subject to censoring because not all individuals will experience the event of interest (disease onset or death) during the study follow-up. Most complex human diseases, including breast cancer, prostate cancer, and bipolar, exhibit

variable ages at onset [Claus et al., 1990; Carter et al., 1992; Stine et al., 1995]. Our work was partly motivated by the Collaborative Study on the Genetics of Alcoholism (COGA) [Begleiter et al., 1995], in which about 54% of individuals were unaffected with alcoholism at the time of interview.

The VC methods have been extended to deal with trait censoring. Epstein et al. [2003] proposed a tobit VC model for left-censored traits, which assumes that the latent trait values within a family follow a multivariate normal distribution, possibly after a known transformation. It is difficult to identify the true transformation, and incorrect transformations can inflate type I error and reduce power. Pankratz et al. [2005] incorporated normal random effects into the Cox [1972] proportional hazards model for right-censored data and estimated the unknown parameters by maximizing the so-called penalized partial likelihood (PPL). This method is not statistically efficient because it disregards some useful information in the observed-data likelihood [Ripatti and Palmgren, 2000], and the Laplace approximation used in computing the PPL function may not be accurate enough to produce unbiased parameter estimates. Both Epstein et al. [2003] and Pankratz et al. [2005] focused on linkage analysis. Li [1999; 2002], Li and Zhong [2002] and Zhong and Li [2004] considered proportional hazards models with gamma frailties. These models induce a restricted form of dependence and are limited to sibships. It should be noted that the proportional hazards assumption often fails, even when the trait pertains to event time.

There exist family-based association tests (FBATs) for age-at-onset data, which explore the correlation between the marker genotype and the phenotype. Horvath et al. [2001] derived the FBAT-Exp by assuming a parametric proportional hazards model with exponentially distributed event times. Lange et al. [2004] proposed the FBAT-logrank and FBAT-Wilcoxon based on the logrank and Wilcoxon statistics. The FBAT-logrank is closely related to the tests of Mokliatchouk et al. [2000], Shih and Whittemore [2002], and Hsu [2003], which were derived under proportional hazards models. The FBATs do not incorporate environmental risk factors or the familial correlation of the trait values and are limited to nuclear families. Thus, the FBATs tend to be less flexible and less powerful than the VC methods.

In this article, we provide robust and powerful VC methods for mapping QTLs with censored trait data in general pedigrees. These methods are derived from a broad class of semiparametric transformation models with random effects. The transformation models include both proportional and non-proportional hazards models; the random effects may consist of major gene effects, polygenic effects, and common environmental effects. We develop efficient likelihood-based estimation and testing procedures under the proposed models.

Our approach is completely general in that it allows linkage and association analyses for any kind of censored traits in extended pedigrees. Indeed, our work unifies and extends substantially the existing methods. In particular, we extend the work of Epstein et al. [2003] to allow an unknown transformation, time-varying environmental factors, and non-normal error distributions, and extend the work of Pankratz et al. [2005], Li [1999; 2002], Li and Zhong [2002], and Zhong and Li [2004] to allow non-proportional hazards models and a rich family of random-effect distributions. We also generalize the FBATs to handle extended pedigrees.

We have implemented the new methods in an efficient and reliable computer program, which is freely available for public use. Extensive simulation studies demonstrate that the proposed methods are considerably more powerful than the existing ones while providing accurate control of the type I error. An application to the aforementioned COGA data is provided.

## METHODS

Suppose that the data contain $n$ families or general pedigrees, with $n_i$ individuals in the $i$th pedigree. We first consider right-censored trait data. Let $T_{ij}$ denote the latent quantitative trait (e.g., age at onset) for the $j$th individual of the $i$th pedigree, and let $C_{ij}$ be the corresponding censoring time. The observation on the trait value consists of two components: $Y_{ij} = \min(T_{ij}, C_{ij})$ and $\Delta_{ij} = I(T_{ij} \le C_{ij})$, where $I(\cdot)$ is the indicator function. Let $\mathbf{X}_{ij}$ be a vector of observed covariates or environmental factors. Consider a (possibly multiallelic) candidate gene coded by $\mathbf{Z}_{ij}$ for the $j$th individual of the $i$th pedigree, which may be a vector incorporating both additive and dominant effects. In the simplest case of a diallic marker (with alleles A and B) with additive genetic effects, $Z_{ij}$ is a scalar and can be coded as $Z_{ij} = -1, 0$, or 1 according to whether this indivi-

dual has genotype B/B, A/B, or A/A, respectively, so that the number of A alleles is $Z_{ij}+1$.

We consider the following class of semiparametric linear transformation models with random effects

$$H(T_{ij}) = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_{ij} + \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{X}_{ij} + R_{ij} + \varepsilon_{ij} \qquad (1)$$

where $H$ is an unknown increasing function, $\boldsymbol{\beta}$ is a set of additive and/or dominant genetic effects, $\boldsymbol{\gamma}$ is a set of fixed covariates effects, $R_{ij}$ is a random effect due to the major gene (after accounting for the marker association) and other genes at unlinked loci, and $\varepsilon_{ij}$ is an individual-specific residual error. In this formulation, association is parameterized by the mean structure whereas linkage is represented by the covariance structure [Fulker et al., 1999; Abecasis et al., 2000; Cardon and Abecasis, 2000]. The choices of the extreme-value and standard logistic distributions for $\varepsilon_{ij}$ correspond to the proportional hazards model [Cox, 1972] and the proportional odds model [Bennett, 1983], respectively. In view of the linear-model form of (1), a more natural choice for $\varepsilon_{ij}$ is the normal distribution.

Write $\mathbf{R}_i = (R_{i1}, \ldots, R_{in_i})^{\mathrm{T}}$. The random effects $\mathbf{R}_i$ represent the within-pedigree correlation of the quantitative traits. The most popular choice for the distribution of $\mathbf{R}_i$ is the multivariate normal distribution with mean zero and variance-covariance matrix $\Sigma_i = \sigma_m^2\Sigma_{mi} + 2\sigma_p^2\Sigma_{pi}$, where $\Sigma_{mi}$ contains the proportions of alleles at the major locus that are IBD among the relative pairs in the $i$th family, $\Sigma_{pi}$ is the matrix of kinship coefficients which depend only on the relatedness of the relative pairs, and $\sigma_m^2$ and $\sigma_p^2$ are the phenotypic variances explained by linkage with the candidate marker and other genes at unlinked loci, respectively.

To avoid detecting spurious association induced by population admixture, we decompose the marker genotype score $\mathbf{Z}_{ij}$ into orthogonal between- and within-family components [Fulker et al., 1999; Abecasis et al., 2000]: $\mathbf{b}_{ij}$ is the expected genotype score conditional on family data, and $\mathbf{w}_{ij}$ is the deviation from this expectation. Let $M_{ij}$ and $F_{ij}$ index the male and female parents of the $j$th individual in the $i$th family. In nuclear families, $\mathbf{b}_{ij}$ is defined as $(\mathbf{Z}_{F_{ij}} + \mathbf{Z}_{M_{ij}})/2$ if parental genotypes are available and as the average of the $\mathbf{Z}_{ij}$ among siblings of the $i$th family otherwise. In general pedigrees, $\mathbf{b}_{ij} = \mathbf{Z}_{ij}$ for genotyped founders; for non-founders, $\mathbf{b}_{ij}$ is $(\mathbf{Z}_{F_{ij}} + \mathbf{Z}_{M_{ij}})/2$ if both $\mathbf{b}_{F_{ij}}$ and $\mathbf{b}_{M_{ij}}$ are defined and is the average genotype score among the full siblings of the $j$th individual in the $i$th pedigree otherwise.

With the above orthogonal decomposition of the genotype scores, we modify model (1) as

$$H(T_{ij}) = \boldsymbol{\beta}_b^{\mathrm{T}}\mathbf{b}_{ij} + \boldsymbol{\beta}_w^{\mathrm{T}}\mathbf{w}_{ij} + \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{X}_{ij} + R_{ij} + \varepsilon_{ij} \qquad (2)$$

where $\boldsymbol{\beta}_b$ and $\boldsymbol{\beta}_w$ pertain to the between-family and within-family effects. This model is a generalization of standard VC models in that it reduces to the model of Abecasis et al. [2000] if the error distribution is normal and the transformation function is known. It is difficult to determine the correct transformation, especially for event times. By leaving the transformation function unspecified, model (2) allows arbitrarily distributed traits while retaining all the attractive features of VC models. For event time data, it is desirable to extend model (2) to accommodate environmental factors or covariates that vary over time. This extension is provided in equation (A.1) of the Appendix.

Let $\boldsymbol{\xi}$ denote the variance parameters $\sigma_m^2$ and $\sigma_p^2$, and let $\boldsymbol{\theta}$ denote the complete set of parameters $\boldsymbol{\beta}_b, \boldsymbol{\beta}_w, \boldsymbol{\gamma}, \boldsymbol{\xi}$, and $H$. The likelihood function for $\boldsymbol{\theta}$ is given in expression (A.2) of the Appendix. The maximum likelihood estimator is denoted by $\hat{\boldsymbol{\theta}}$. Because the likelihood is a complex function involving the non-parametric function $H(\cdot)$, the calculation of $\hat{\boldsymbol{\theta}}$ is not a trivial matter. We describe some efficient numerical methods in the Appendix. The maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ is approximately unbiased, normally distributed, and statistically efficient. These results imply that the unknown transformation $H(\cdot)$ can be correctly estimated from the data and the likelihood-based test statistics are the most powerful among all valid test statistics.

Although the proportional hazards model (i.e., the extreme-value error distribution) is commonly used in the analysis of uncorrelated event time data, the normal error distribution is a more natural choice for model (2) with normal random effects, especially when the quantitative trait does not pertain to event time. It is also computationally simpler to use the normal error distribution. Thus, all the numerical results in the sequel pertain to the normal error model. Model (2) with normal random effects and normal error is reminiscent of standard VC models [Amos, 1994; Abecasis et al., 2000], but is considerably less restrictive because the transformation function $H$ is unspecified.

For left-censored trait data, we simply regard $-T_{ij}$ and $-C_{ij}$ as the quantitative trait value and

censoring value, respectively. Then all the above results hold. Epstein et al. [2003] presented a tobit VC model to account for left censoring. Their model takes a similar form to our model (1); however, it assumes that the transformation function $H$ is known and does not model marker association. As indicated earlier, it is desirable to leave the transformation function unspecified, especially for event times.

We can perform various hypothesis testing under model (2). For the linkage analysis, we omit the association components in (2) and test the null hypothesis $H_0: \sigma_m^2 = 0$ against the alternative $H_A: \sigma_m^2 > 0$. We can assess whether there is association between the candidate marker and quantitative trait by testing the null hypothesis $H_0: \boldsymbol{\beta}_w = \mathbf{0}$. We can also test for the presence of population admixture, $H_0: \boldsymbol{\beta}_b = \boldsymbol{\beta}_w$. If no population admixture is detected, we can use the more powerful test based on model (1). For each hypothesis test, we calculate the likelihood ratio statistic

$$\mathrm{LR} = -2[\log L(\widetilde{\boldsymbol{\theta}}) - \log L(\widehat{\boldsymbol{\theta}})]$$

where $\widetilde{\boldsymbol{\theta}}$ is the restricted maximum likelihood estimator of $\boldsymbol{\theta}$ under the null hypothesis. For testing association, LR is approximately $\chi^2$ distributed with the degrees of freedom being the dimension of $\boldsymbol{\beta}_w$. In the special case of additive genetic effects, the null distribution of LR is approximately $\chi_1^2$ at a diallelic marker locus. For testing the variance parameters, the distribution of LR is approximated by a mixture of $\chi^2$ distributions [Self and Liang, 1987].

# RESULTS

## SIMULATION STUDIES

We conducted extensive simulation studies to assess the performance of the new methods and to compare them with the best existing methods. We assumed an additive QTL, Q, with two alleles $Q_1$ and $Q_2$ and simulated a diallelic marker M with alleles $M_1$ and $M_2$. We created population admixture by mixing in equal proportions families from two populations (A and B) with different QTL and marker allele frequencies: in population A, $p_{Q_1} = p_{M_1} = 0.25$; in population B, $p_{Q_1} = p_{M_1} = 0.75$. For each simulation set-up, we generated 10,000 data sets, each with 100 nuclear families. The number of siblings in a family was set to 2, 3, 4, and 5 with probabilities 0.3, 0.3, 0.2, and 0.2, respectively. The parental genotypes were assumed to be known.

We first considered linkage analysis with left-censored data. We generated trait values from the model

$$H(T_{ij}) = \beta Z_{ij} + \gamma_0 + \gamma_1 X_{1ij} + \gamma_2 X_{2ij} + g_{ij} + e_{ij} \quad (3)$$

where $H(t) = (25 \log t)^{2/3} \mathrm{sign}(\log t)$, $\gamma_0 = -1.5$, $\gamma_1 = -1.0$, $\gamma_2 = 1.0$, $Z_{ij}$ is the QTL genotype score, $X_{1ij}$ is a binary variable with 0.5 probability of being 1, $X_{2ij}$ is an exponential variable with mean value of 2 shifted to the right by $|Z_{ij}| - 0.5$, and $g_{ij}$ and $e_{ij}$ are independent zero-mean normal variables with variances $\sigma_p^2$ and $\sigma_e^2$. (The function sign($x$) takes value 1 if $x$ is non-negative and $-1$ otherwise.) We chose the variance parameters to yield different levels of overall genetic heritability $h^2 = (\sigma_m^2 + \sigma_p^2)/(\sigma_m^2 + \sigma_p^2 + \sigma_e^2)$ and major gene heritability $h_m^2 = \sigma_m^2/(\sigma_m^2 + \sigma_p^2 + \sigma_e^2)$. In particular, we considered the following six scenarios:

| Model | $\sigma_m^2$ | $\sigma_p^2$ | $\sigma_e^2$ | $h_m^2$ | $h^2$ |
|---|---|---|---|---|---|
| a | 0.0 | 1.6 | 0.4 | 0.0 | 0.8 |
| b | 0.4 | 1.2 | 0.4 | 0.2 | 0.8 |
| c | 0.8 | 0.8 | 0.4 | 0.4 | 0.8 |
| d | 0.0 | 1.2 | 0.8 | 0.0 | 0.6 |
| e | 0.4 | 0.8 | 0.8 | 0.2 | 0.6 |
| f | 0.8 | 0.4 | 0.8 | 0.4 | 0.6 |

Scenarios a and d pertain to the null hypothesis, and the others to alternative hypotheses unless the recombination fraction between the QTL and the marker locus is 0.5. We set $\beta = 1.633\sigma_m$. Figure 1 shows the distribution of the trait values for the first simulated data set under scenario a. This type of distribution is commonly seen in real studies. After simulating the trait values, we censored those values below the 25th percentile of the trait distribution.

We evaluated the proposed semiparametric linkage test for $H_0: \sigma_m^2 = 0$ as well as the parametric test of Epstein et al. [2003]. For the latter, we performed various transformations, including the true transformation, Box-Cox transformation, and normal-score transformation. The parametric test with the true transformation is an idealized situation in which the normality assumption holds after a known transformation.

The results of these studies are presented in Table I and Figure 2. Table I shows the type I error and power at the nominal significance level of 5% at the true QTL, while Figure 2 displays the results under scenario f with the recombination fraction

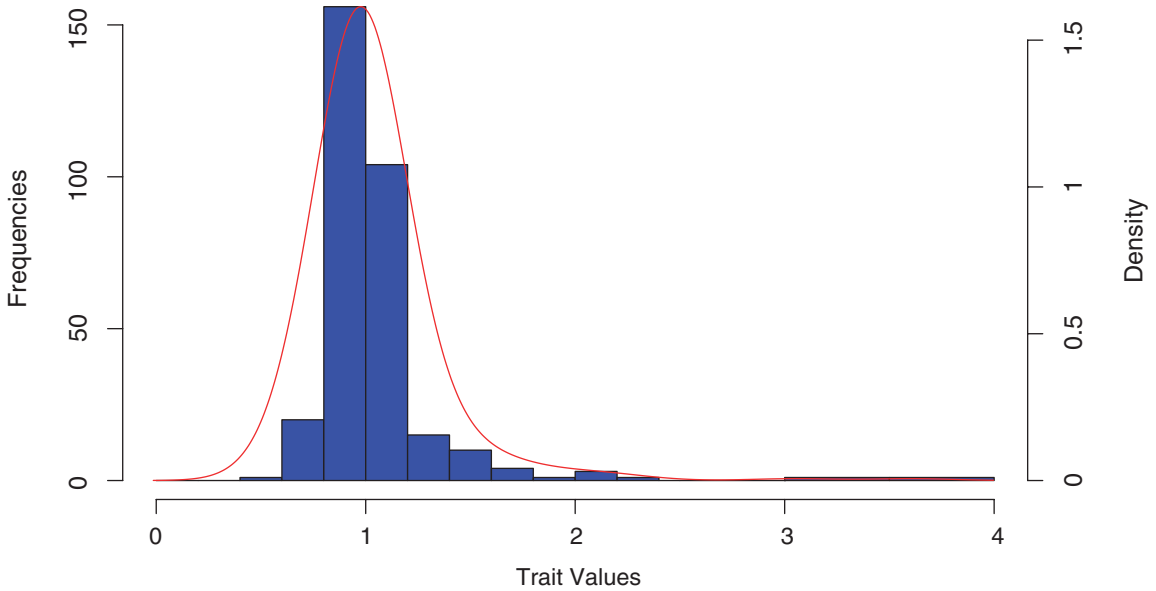


**Fig. 1. The distribution of the latent trait values for a simulated data set: the histogram is shown in blue bars and estimated density function in red line.**

**TABLE I. Type I error and power (%) of the linkage tests at the nominal significance level of 5% for left-censored traits**

| | | Epstein et al. method with transformation | | | |
|---|---|---|---|---|---|
| Model | New method | True | None | Box-Cox | Normal score |
| a | 5.48 | 5.95 | 17.08 | 9.30 | 8.24 |
| b | 53.54 | 55.12 | 36.54 | 44.52 | 37.34 |
| c | 94.68 | 96.10 | 52.31 | 80.66 | 75.02 |
| d | 4.70 | 4.95 | 13.81 | 6.64 | 7.55 |
| e | 45.74 | 46.60 | 34.02 | 37.98 | 32.61 |
| f | 89.63 | 91.23 | 50.14 | 75.04 | 67.60 |

*Note*: Models a and d pertain to the null hypothesis and the others to alternative hypotheses.

between the marker locus and the QTL ranging from 0 to 0.5. The new method provides accurate control of the type I error in all cases and has virtually the same power as the Epstein et al. method with the true transformation. Without transformation, the type I error of the Epstein et al. method is very off. With the Box-Cox and normal-score transformations, the type I error is still inflated. Although it has smaller type I error than the Epstein et al. method with incorrect transformations, the new method is more powerful. Under scenario f, the power of the new linkage test (at the true QTL) is 89.6% at the 5% nominal significance level, as compared to 50.1%, 75.0%, and 67.6% for the Epstein et al. tests without transformation, with Box-Cox and normal-score transformations, respectively.

Our second set of simulation studies was concerned with the association analysis of left-censored data. We considered the same model as in the above linkage studies except that different values for $\sigma_m^2, \sigma_p^2$, and $\sigma_e^2$ were used. We introduced linkage disequilibrium (LD) within each population between the QTL and marker locus in the parental chromosomes. In each population, LD is measured by $D = p_{M_1Q_1} - p_{M_1}p_{Q_1}$, where $p_{M_1Q_1}$ is the frequency of haplotype $M_1Q_1$. The standardized LD coefficient is $D' = D/D_{\max}$, where $D_{\max} = \min(p_{M_1}, p_{Q_1}) - p_{M_1Q_1}$. When there is no LD in either population, LD exists in the pooled population with $D' = 0.25$. The marker locus is tightly linked to the QTL with a recombination fraction of 0, but we considered different levels of $D'$. We set $\sigma_m^2, \sigma_p^2$, and $\sigma_e^2$ to 0.16, 0.64, and 1.2, corresponding to the overall heritability $h^2$ of 0.4 and the major gene heritability $h_m^2$ of 0.08. The value of $\beta$ became 0.653.

We assessed the performance of the proposed semiparametric association test for $H_0: \beta_w = 0$ at the nominal significance level of 5% and compared it with the parametric test. The latter is derived from model (2) with a specified transformation function. The results of these studies are presented in Table II. The semiparametric method performs nearly as well as its parametric counterpart with the true transformation.

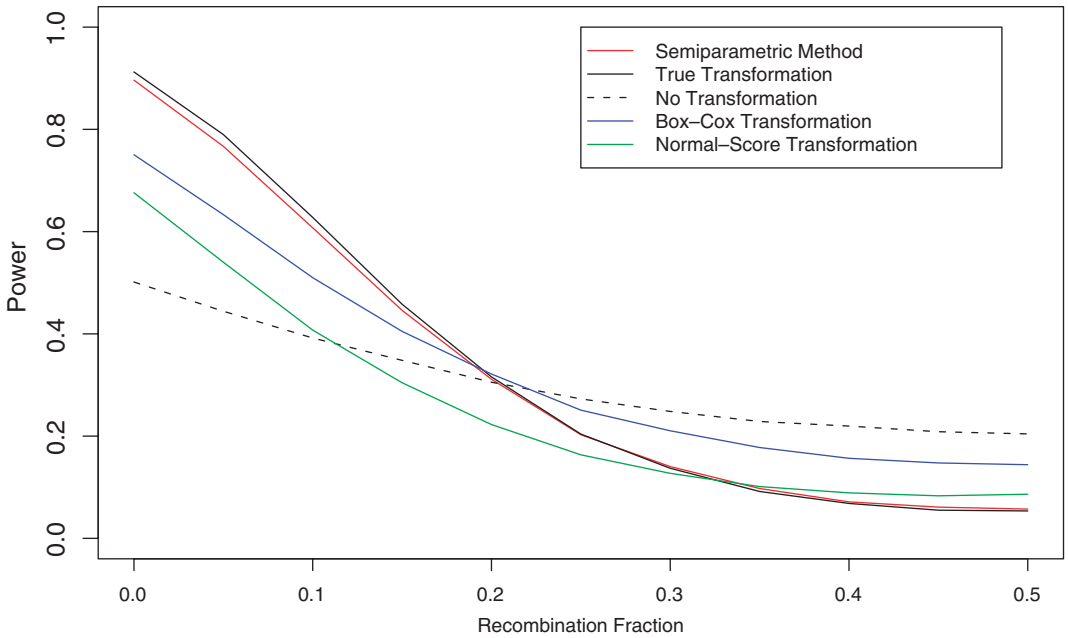


**Fig. 2. Type I error and power of the semiparametric linkage test versus the Epstein et al. [2003] test with various transformations at the nominal significance level of 0.05 for left-censored traits.**

**TABLE II. Type I error and power (%) of the association tests at the nominal significance level of 5% for left-censored traits**

| | | Parametric method with transformation | | | |
|---|---|---|---|---|---|
| $D'$ | New method | True | None | Box-Cox | Normal score |
| 0.00 | 4.94 | 5.28 | 7.94 | 7.07 | 5.73 |
| 0.25 | 13.77 | 14.83 | 16.85 | 17.09 | 7.17 |
| 0.50 | 41.68 | 43.10 | 35.27 | 39.46 | 23.77 |
| 0.75 | 75.18 | 76.66 | 57.03 | 66.02 | 51.68 |
| 1.00 | 94.73 | 95.34 | 76.28 | 86.63 | 80.27 |

With incorrect transformations, the type I error of the parametric test is inflated and the power is reduced.

The additive effect of the marker on the phenotype is $\eta = \beta D/p_{M_1}p_{M_2}$ [Cardon and Abecasis, 2000]. The results for the estimation of this parameter are summarized in Table III. The proposed estimator appears to be unbiased. The standard error estimator reflects accurately the true variation, and the confidence intervals have proper coverage probabilities. As expected, the effect size of the marker decreases as the LD between the QTL and marker alleles becomes weaker.

Our third set of studies was concerned with right-censored age-at-onset data. We generated the ages-at-onset and censoring times from the following models:

$$\log \Lambda(T_{ij}) = \beta Z_{ij} + \gamma_1 X_{1ij} + \gamma_2 X_{2ij} + g_{ij} + e_{ij} \quad (4)$$

and

$$\log \Lambda(C_{ij}) = e_{ij} \quad (5)$$

where $\Lambda(t) = t^{\tau+1}/b^{\tau}(\tau+1)$ with $b = 72$ and $\tau = 5$, $\gamma_1 = -1.0$, $\gamma_2 = 1.0$, $Z_{ij}$ is the QTL genotype score, $X_{1ij}$ is a binary variable with 0.5 probability of being 1, $X_{2ij}$ is an independent standard normal variable, and $g_{ij}$ is a zero-mean normal polygenic random effect with variance $\sigma_p^2$. We set $\sigma_m^2$ and $\sigma_p^2$ to 0.2 and 0.80, so that $\beta = 0.73$. We considered the extreme-value and normal distributions for $e_{ij}$. The mean and variance of the normal error distribution were chosen so that the two error distributions have the same mean and variance. The censoring rate was approximately 50%. We used the proposed method to test the null hypothesis $H_0: \beta_w = 0$ and compared it to the FBAT-logrank, FBAT-Wilcoxon, and FBAT-Exp.

The results of these studies are presented in Table IV. Although the FBATs have reasonable type I error, the new method is more powerful than the FBATs while providing accurate control of the type I error even when the error distribution is mis-specified. In the case of complete LD (i.e., $D' = 1$) and extreme-value error distribution, the power of the proposed test is 64.0% at the nominal

significance level of 1%, as compared to 40.0%, 40.2%, and 51.3% for the FBAT-logrank, FBAT-Wilcoxon, and FBAT-Exp, respectively.

Finally, we generated data from models (4) and (5) but sampled only those sibships with one or more affected individuals. Table V presents the results with such ascertained families. Although not developed to correct for ascertainment, the proposed method seems to be robust to ascertainment bias: it continues to provide accurate control of the type I error and is more powerful than the FBATs. As expected, sampling families with one or more affected individuals is more efficient than the population-based sampling. The power improvement will be more appreciable for rare diseases.

**TABLE III. Summary statistics for the estimation of the marker effects with left-censored traits**

| $D'$ | True value | Mean | SE | SEE | CP(%) |
|---|---|---|---|---|---|
| 0.00 | 0.000 | 0.003 | 0.172 | 0.169 | 94.7 |
| 0.25 | 0.149 | 0.148 | 0.170 | 0.167 | 94.6 |
| 0.50 | 0.298 | 0.291 | 0.169 | 0.165 | 94.3 |
| 0.75 | 0.447 | 0.433 | 0.168 | 0.162 | 94.3 |
| 1.00 | 0.596 | 0.578 | 0.167 | 0.159 | 93.8 |

*Note*: The original value of $\eta$ is divided by $\sigma_e=\sqrt{1.2}$, so that $\eta$ and $\widehat{\beta}_w$ are compared on the same scale. Mean and SE are the sampling mean and sampling standard error of the parameter estimator; SEE is the mean of the standard error estimator, and CP is the coverage probability of the 95% confidence interval.

## COGA STUDY

Alcoholism is a disease that tends to run in families and results in part from genetic risk factors. COGA is a nine-site national collaboration with the goal of identifying and characterizing those genetic factors that affect the susceptibility to alcohol dependence and related phenotypes [Begleiter et al., 1995]. Initial ascertainment of COGA probands was performed by screening consecutive admissions at treatment facilities. To be recruited into the COGA study, probands had to meet both the diagnostic criteria for alcohol dependence by the DSM-III-R standards [American Psychiatric Association, 1987] and the criteria for definite alcoholism specified by Feighner et al. [1972]. Thus, the COGA sample is representative of a severely alcohol-dependent population. A subset of COGA families with at least three alcohol-dependent first-degree relatives was

**TABLE IV. Type I error and power (%) of the association tests at the nominal significance level of 5% for right-censored traits**

| | Extreme-value error | | | | Normal error | | | |
|---|---|---|---|---|---|---|---|---|
| $D'$ | New | FBAT-L | FBAT-W | FBAT-E | New | FBAT-L | FBAT-W | FBAT-E |
| 0.00 | 4.79 | 4.64 | 4.69 | 4.79 | 4.62 | 4.91 | 4.59 | 4.75 |
| 0.25 | 10.66 | 9.03 | 9.06 | 9.90 | 11.82 | 8.45 | 8.99 | 9.63 |
| 0.50 | 30.02 | 21.62 | 21.69 | 25.53 | 33.29 | 21.44 | 22.56 | 25.21 |
| 0.75 | 58.56 | 42.51 | 42.76 | 51.15 | 63.36 | 42.47 | 44.94 | 50.27 |
| 1.00 | 84.02 | 66.61 | 66.77 | 75.94 | 87.26 | 66.31 | 69.60 | 75.85 |

*Note*: FBAT-L, FBAT-W, and FBAT-E are the FBAT-logrank and FBAT-Wilcoxon due to Lange et al. [2004], and the FBAT-Exp due to Horvath et al. [2001], respectively.

**TABLE V. Type I error and power (%) of the association tests at the nominal significance level of 5% for right-censored traits in ascertained families**

| | Extreme-value error | | | | Normal error | | | |
|---|---|---|---|---|---|---|---|---|
| $D'$ | New | FBAT-L | FBAT-W | FBAT-E | New | FBAT-L | FBAT-W | FBAT-E |
| 0.00 | 5.18 | 4.98 | 5.00 | 5.30 | 5.58 | 5.32 | 5.08 | 5.11 |
| 0.25 | 11.70 | 9.28 | 9.38 | 10.47 | 12.30 | 9.18 | 9.52 | 10.32 |
| 0.50 | 31.22 | 22.09 | 21.72 | 27.49 | 34.59 | 22.07 | 23.09 | 27.30 |
| 0.75 | 62.42 | 45.50 | 45.59 | 55.83 | 66.25 | 43.69 | 46.82 | 54.79 |
| 1.00 | 86.25 | 68.42 | 69.03 | 79.73 | 89.43 | 67.87 | 70.72 | 79.01 |

*Note*: FBAT-L, FBAT-W, and FBAT-E are the FBAT-logrank and FBAT-Wilcoxon due to Lange et al. [2004], and the FBAT-Exp due to Horvath et al. [2001], respectively.

identified as suitable for a genetic linkage study. With non-genotyped individuals included for linking in the pedigrees, the families selected for genotyping as part of Genetic Analysis Workshop 14 had a total of 1,614 individuals. Family sizes ranged from 5 to 32.

We considered the age at onset of ALDX1, the DSM-III-R+Feighner classification status for alcohol dependence. The ages at interview were the censoring times for the unaffected individuals. Among the 1,614 individuals in the study, 643 were affected with ALDX1, 626 of whom had known ages at onset. The final data set for our analysis consisted of 1,371 individuals, including 626 affected individuals and 745 unaffected individuals. Of the 626 affected individuals, 424 were males, as opposed to 229 males in the unaffected individuals.

Figure 3 presents the Kaplan-Meier estimates of the disease-free probabilities for the onset of ALDX1. The age-at-onset distributions resemble Figure 1, with most events occurring between ages 15 and 40. It would be difficult to achieve approximate normality for such distributions through simple transformations. Preliminary analysis revealed that gender was associated with the age at onset of ALDX1; males developed disease earlier than females. Previous linkage analysis showed a linked region on chromosome 14 [Palmer et al., 1999]. We performed association analysis under model (2) using 172 SNPs on chromosome 14 from Illumina and included gender as a covariate in the model. We also considered the FBAT-logrank, FBAT-Wilcoxon, and FBAT-Exp. Since the FBATs require nuclear families whereas the COGA data have extended pedigrees, we generalized the FBATs by using the PDT idea of Monks and Kaplan [2000]. The resulting tests are termed the PDT-logrank, PDT-Wilcoxon, and PDT-Exp. The FBATs and the PDTs differ in the calculation of the variance for the test statistic, although they are asymptotically equivalent for nuclear families.

Figure 4 displays the LOD scores of the VC method and the three PDTs. The LOD scores were obtained by dividing the original LR statistics by 2log10. The LOD score curves from the VC method and the PDT-logrank reached their peaks at the same location of 0 cM for SNP rs1972373, with peak values 2.67 and 2.25, respectively. The LOD score curve from the PDT-Wilcoxon reached its peak at the location of 41.7 cM for SNP rs944044 with peak value of 1.56, whereas the LOD score curve from the PDT-Exp reached its peak at the location of 0.6 cM for SNP rs1057605 with peak value of 2.43. The corresponding $p$-values are $4.55 \times 10^{-4}$, $1.29 \times 10^{-3}$, $7.40 \times 10^{-3}$, $8.22 \times 10^{-4}$, for the VC method, PDT-logrank, PDT-Wilcoxon, and PDT-Exp, respectively. With the Bonferroni correction for multiple testing, only the VC

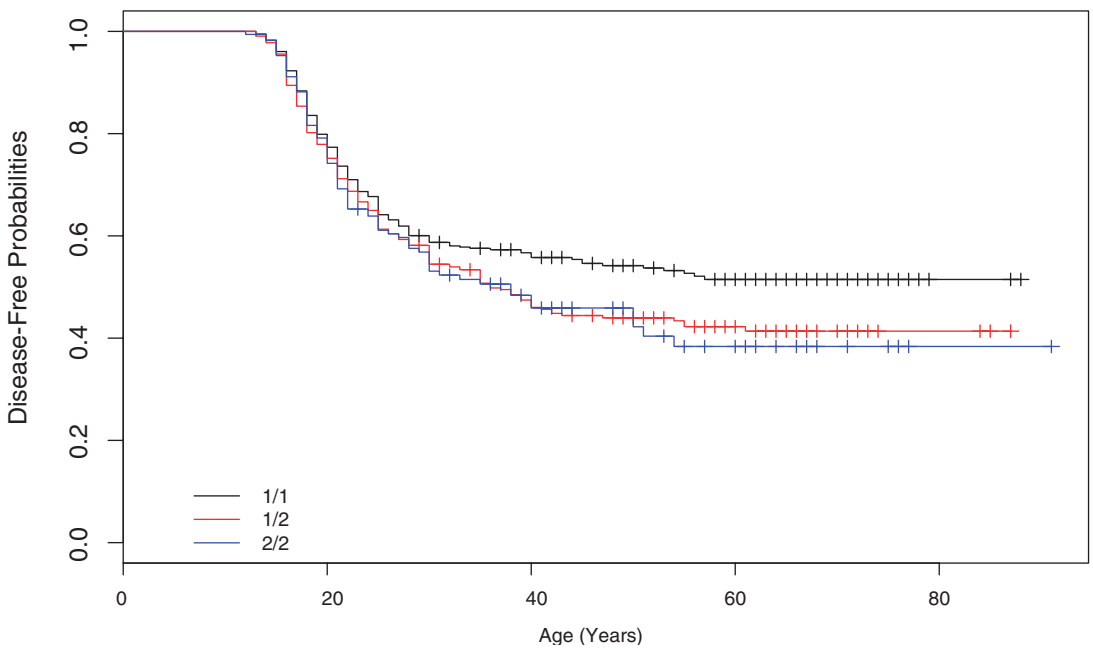


**Fig. 3. Kaplan-Meier estimates of the disease-free probabilities stratified by the genotype of SNP rs1972373 for the age at onset of ALDX1 in the COGA study.**

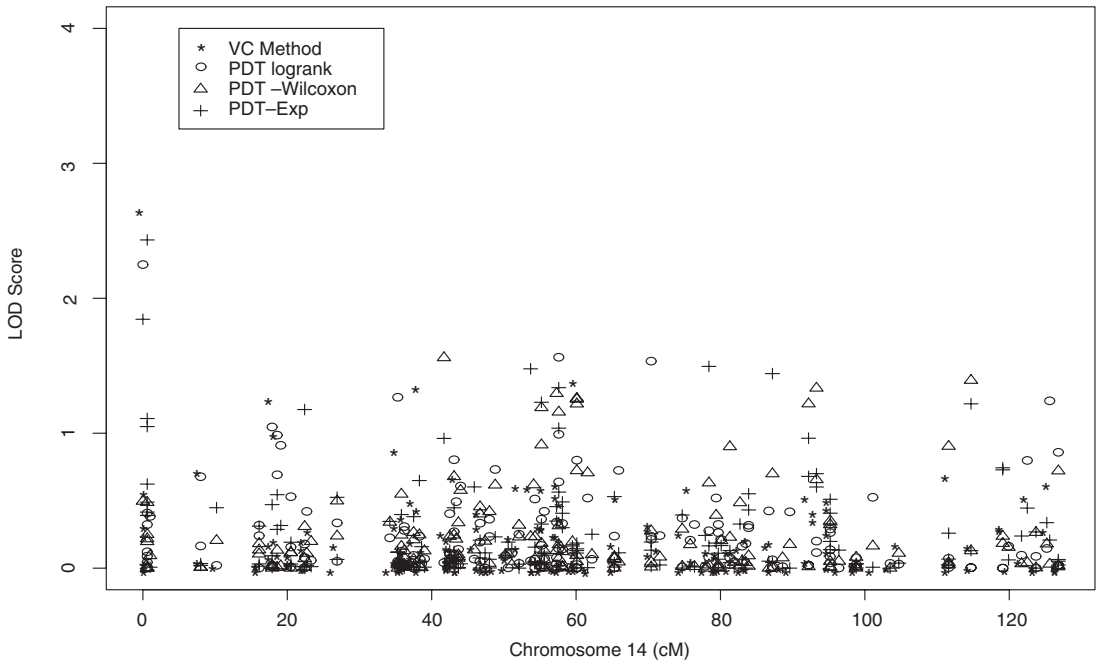


**Fig. 4. LOD scores based on the semiparametric VC method and the three PDT methods for the age at onset of ALDX1 on chromosome 14 in the COGA study.**

method yielded significant association between the SNP marker and the onset age of ALDX1 at the nominal significance level of 0.1, with an adjusted $p$-value of 0.078.

Since the results are only marginally significant and the truth is unknown, no firm conclusions can be drawn regarding the genetic association or the comparative value of different approaches. Nevertheless, this example shows that the proposed method can handle general pedigrees and can yield meaningfully different results than the existing methods.

## DISCUSSION

QTL mapping with censored data is an important and challenging problem. The existing solutions have significant limitations. In this article, we provide a general approach that unifies and extends the existing literature. Our approach is applicable to left- or right-censored traits, accommodates arbitrary trait distributions, allows extended pedigrees with missing genotype data, and provides a common framework for linkage and association analyses. As demonstrated by the simulation studies, the new methods can greatly increase the power of QTL mapping over the best existing methods.

We have implemented the new methods in a cost-free computer program (D.Y.L.'s Web site: http://www.bios.unc.edu/∼lin). The computing time depends mainly on the number of censored observations in a family. For the COGA data, the largest number of censored observations in a family is 20, in which case it took less than 4 min on an IBM BladeCenter HS-20 machine to perform the association analysis at one locus. With sibship sizes ranging from 2 to 5 in the simulation studies, the analysis at one position took only 3 sec.

This article is a substantial generalization of our earlier work on QTL mapping with non-censored data. Diao and Lin [2005] extended the traditional VC model for linkage analysis [Amos, 1994] by allowing an arbitrary transformation, and Diao and Lin [2006] extended this model further to association mapping. The semiparametric transformation models proposed in this article are more general than our previous models in that they allow time-varying covariates and non-normal error distributions. As is evident in the Appendix, trait censoring poses considerable new challenges, both theoretically and computationally. Simulation studies revealed that applying our previous methods to censored trait data would have detrimental effects on the type I error and power (data not shown).

In some studies, families are selected on the basis of the trait values of their members. Li and Zhong [2002] and Zhong and Li [2004] proposed a retrospective likelihood approach to correct for ascertainment bias, which requires that the disease prevalence is known or can be estimated externally. de Andrade and Amos [2000] proposed two ascertainment correction methods: one is to divide the likelihood by the probability that the proband falls into the specified ascertainment region, and the other is to condition on the observed trait values. They showed that the power to detect linkage is similar regardless of whether the data are corrected for ascertainment. Our simulation results revealed that, even without ascertainment correction, the proposed association tests have reasonable type I error and are more powerful than the existing tests. Further investigation is warranted.

In association studies, one often examines a large number of SNPs in a chromosomal region. To guard against an abundance of false-positive results, one needs to adjust for multiple testing. The commonly adopted Bonferroni correction is conservative because the test statistics for SNPs in LD are correlated. Recently, Lin [2005] proposed a Monte Carlo procedure that properly accounts for the correlatedness of polymorphism data. It would be worthwhile to apply that approach to the proposed association tests.

## ACKNOWLEDGMENTS

The authors are grateful to the COGA investigators and Jean W. MacCluer for providing the COGA data from GAW14, which was supported in part by the NIH grant GM31575.

## REFERENCES

Abecasis GR, Cardon LR, Cookson WOC. 2000. A general test of association for quantitative traits in nuclear families. Am J Hum Genet 66:279–292.

Almasy L, Blangero J. 1998. Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet 62:1198–1211.

American Psychiatric Association. 1987. Diagnostic and Statistical Manual of Mental Disorders, 3rd edition. Washington, DC.

Amos CI. 1994. Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet 54: 535–543.

Amos CI, Zhu DK, Boerwinkle E. 1996. Assessing genetic linkage and association with robust components of variance approaches. Ann Hum Genet 60:143–160.

Begleiter H, Reich T, Hesselbrock V, Porjesz B, Li TK, Schuckit MA, Edenberg HJ, Rice JP. 1995. The collaborative study on the genetics of alcoholism. Alcohol Health Res World 19:228–236.

Bennett S. 1983. Analysis of survival data by proportional odds model. Stat Med 2:273–277.

Bickel PJ, Klassen CAJ, Ritov Y, Wellner JA. 1993. Efficient and Adaptive Estimation in Semiparametric Models, Baltimore: Johns Hopkins University Press.

Cardon LR, Abecasis GR. 2000. Some properties of a variance components model for fine-mapping quantitative trait loci. Behav Genet 30:235–243.

Carter BS, Beaty HB, Steinberg GD, Childs B, Walsh PC. 1992. Mendelian inheritance of familial prostate cancer. Proc Natl Acad Sci USA 89:3367–3371.

Claus EB, Risch NJ, Thompson WD. 1990. Using age of onset to distinguish between subforms of breast cancer. Ann Hum Genet 54:169–177.

Cox DR. 1972. Regression models and life tables (with discussion). J R Stat Soc B 34:187–220.

de Andrade M, Amos CI. 2000. Ascertainment issues in variance components models. Genet Epidemiol 19:333–344.

Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B 39: 1–38.

Diao G, Lin DY. 2005. A powerful and robust method for mapping quantitative trait loci in general pedigrees. Am J Hum Genet 77:97–111.

Diao G, Lin DY. 2006. Improving the power of association tests for quantitative traits in family studies. Genet Epidemiol 30: 301–313.

Epstein MP, Lin XH, Boehnke M. 2003. A Tobit variance-component method for linkage analysis of censored trait data. Am J Hum Genet 72:611–620.

Feighner JP, Robins E, Guze SB et al. 1972. Diagnostic criteria for use in psychiatric research. Arch Gen Psychiatry 26:57–63.

Fulker DW, Cherny SS, Sham PC, Hewitt JK. 1999. Combined linkage and association sib-pair analysis for quantitative traits. Am J Hum Genet 64:259–267.

Genz A. 1992. Numerical computation of multivariate normal probabilities. J Comput Graph Stat 1:141–149.

Higgins M, Province M, Heiss G, Eckfeldt J, Ellison RC, Folsom AR, Rao DC, Sprafka M, Williams R. 1996. NHLBI Family Heart Study: objectives and design. Am J Epidemiol 143:1219–1228.

Horvath S, Xu X, Laird NM. 2001. The family based association test method: strategies for studying general genotype-phenotype associations. Eur J Hum Genet 9:301–306.

Hsu L. 2003. Genetic association tests with age at onset. Genet Epidemiol 24:118–127.

Lange C, Blacker D, Laird NM. 2004. Family-based association tests for survival and times-to-onset analysis. Stat Med 23: 179–189.

Li H. 1999. The additive genetic gamma frailty model for linkage analysis. Ann Hum Genet 63:455–468.

Li H. 2002. The additive genetic gamma frailty model for linkage analysis of diseases with variable age of onset using nuclear families. Lifetime Data Anal 8:315–334.

Li H, Zhong X. 2002. Multivariate survival models induced by genetic frailties, with application to linkage analysis. Biostatistics 3:57–75.

Lin DY. 2005. An efficient Monte Carlo approach to assessing statistical significance in genomic studies. Bioinformatics 21: 781–787.

Mokliatchouk O, Blacker D, Rabinowitz D. 2000. Association tests for traits with variable age at onset. Hum Hered 51:46–53.

Monks SA, Kaplan NL. 2000. Removing the sampling restrictions from family-based tests of association for a quantitative-trait locus. Am J Hum Genet 66:576–592.

Palmer LJ, Katrina JT, Burton PR. 1999. Genome-wide linkage analysis using genetic variance components of alcohol dependency-associated censored and continuous traits. Genet Epdemiol 17(Suppl. 1):S283–S288.

Pankratz VS, de Andrade M, Therneau TM. 2005. Random-effect Cox proportional hazards model: general variance components methods for time-to-event data. Genet Epidemiol 28:97–109.

Pinheiro JC, Bates DM. 1995. Approximations to the log-likelihood function in the nonlinear mixed-effects model. J Comput Graph Stat 4:12–35.

Press WH, Teukolsky SA, Vetterling WT, Flannery BP. 1992. Numerical Recipes in C: The Art of Scientific Computing, 2nd edition. New York: Cambridge University Press.

Ripatti S, Palmgren J. 2000. Estimation of multivariate frailty models using penalized partial likelihood. Biometrics 56: 1016–1022.

Self SG, Liang KL. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J Am Stat Assoc 82:605–610.

Shih MC, Whittemore AS. 2002. Tests for genetic association using family data. Genet Epidemiol 22:128–145.

Stine OC, Xu J, Koskela R, McMahon FJ, Gschwend M, Friddle C, Clark CD, McInnis MG, Simpson SG, Breschel TS et al. 1995. Evidence for linkage of bipolar disorder to chromosome 18 with a parent-of-origin effect. Am J Hum Genet 57:1384–1394.

Valle T, Tuomilehto J, Bergman RN, Ghosh S, Hauser ER, Eriksson J, Nylund SJ, Kohtamaki K, Toivanen L, Vidgren G, Tuomilehto-Wolf E, Ehnholm C, Blaschak J, Langefeld CD, Watanabe RM, Magnuson V, Ally DS, Hagopian WA, Ross E, Buchanan TA, Collins F, Boehnke M. 1998. Mapping genes for NIDDM: design of the Finland-United States Investigation of NIDDM Genetics (FUSION) study. Diab Care 21:949–958.

Zeng D, Lin DY, Lin X. 2005. Semiparametric transformation models with random effects for clustered failure time data. Technical Report, Department of Biostatistics, University of North Carolina at Chapel Hill.

Zhong X, Li H. 2004. Score tests of genetic association in the presence of linkage based on the additive genetic gamma frailty model. Biostatistics 5:307–327.

## APPENDIX

To accommodate time-varying covariates, we extend model (2) through the distribution function of $T_{ij}$

$$P(T_{ij} \le t|\mathbf{X}_{ij}, \mathbf{b}_{ij}, \mathbf{w}_{ij}, R_{ij}) = G\left(\int_0^t \mathrm{e}^{-\boldsymbol{\beta}_b^{\mathrm{T}}\mathbf{b}_{ij}-\boldsymbol{\beta}_w^{\mathrm{T}}\mathbf{w}_{ij}-\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{X}_{ij}(s)-R_{ij}}\mathrm{d}\Lambda(s)\right) \quad (A.1)$$

where $\mathbf{X}_{ij}(t)$ pertains to the value of $\mathbf{X}_{ij}$ at time $t$, $G$ is the distribution function of $\mathrm{e}^{\varepsilon_{ij}}$, and $\Lambda(t) = \mathrm{e}^{H(t)}$. Under $G(x) = 1-\mathrm{e}^{-x}$, (A.1) corresponds to the proportional hazards model with random effects and $\Lambda$ is called the cumulative baseline hazard function. In the absence of time-varying covariates, equation (A.1) is equivalent to equation (2). The representation given in equation (A.1) shows that leaving the transformation function in equation (2) unspecified is tantamount to allowing an arbitrary distribution of the trait values.

Under model (A.1), the likelihood function for the set of unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}_b, \boldsymbol{\beta}_w, \boldsymbol{\gamma}, \boldsymbol{\xi}, \Lambda)$ is proportional to

$$\prod_{i=1}^{n}\left[\int_{\mathbf{R}_i}\prod_{j=1}^{n_i}\left\{G'\left(\int_0^{Y_{ij}} \mathrm{e}^{-\boldsymbol{\beta}_b^{\mathrm{T}}\mathbf{b}_{ij}-\boldsymbol{\beta}_w^{\mathrm{T}}\mathbf{w}_{ij}-\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{X}_{ij}(s)-R_{ij}}\mathrm{d}\Lambda(s)\right)\right.\right.$$
$$\left.\times e^{-\boldsymbol{\beta}_b^{\mathrm{T}}\mathbf{b}_{ij}-\boldsymbol{\beta}_w^{\mathrm{T}}\mathbf{w}_{ij}-\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{X}_{ij}(s)-R_{ij}}\lambda(Y_{ij})\right\}^{\Delta_{ij}}$$
$$\times\left\{1-G\left(\int_0^{Y_{ij}} e^{-\boldsymbol{\beta}_b^{\mathrm{T}}\mathbf{b}_{ij}-\boldsymbol{\beta}_w^{\mathrm{T}}\mathbf{w}_{ij}-\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{X}_{ij}(s)-R_{ij}}\mathrm{d}\Lambda(s)\right)\right\}^{1-\Delta_{ij}}$$
$$\left.\times\,\phi_i(\mathbf{R}_i;\boldsymbol{\xi})\mathrm{d}\mathbf{R}_i\right] \quad (A.2)$$

where $\lambda(t)$ is the derivative of $\Lambda(t)$, $G'(x)$ is the derivative of $G(x)$, and $\phi_i(\mathbf{R}_i;\boldsymbol{\xi})$ is the density function of the random effects $\mathbf{R}_i$ indexed by the variance parameters $\boldsymbol{\xi}$. This is a non-parametric likelihood in that $\Lambda(\cdot)$ is a completely arbitrary function. It would be natural to estimate $\boldsymbol{\theta}$ by maximizing (A.2). The maximum of this function does not exist if $\Lambda$ is restricted to be absolutely continuous. Thus, we regard $\Lambda$ as a right-continuous function and maximize the following function:

$$\prod_{i=1}^{n}\left[\int_{\mathbf{R}_i}\prod_{j=1}^{n_i}\left\{G'\left(\int_0^{Y_{ij}} \mathrm{e}^{-\boldsymbol{\beta}_b^{\mathrm{T}}\mathbf{b}_{ij}-\boldsymbol{\beta}_w^{\mathrm{T}}\mathbf{w}_{ij}-\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{X}_{ij}(s)-R_{ij}}\mathrm{d}\Lambda(s)\right)\right.\right.$$
$$\left.\times e^{-\boldsymbol{\beta}_b^{\mathrm{T}}\mathbf{b}_{ij}-\boldsymbol{\beta}_w^{\mathrm{T}}\mathbf{w}_{ij}-\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{X}_{ij}(s)-R_{ij}}\Lambda\{Y_{ij}\}\right\}^{\Delta_{ij}}$$
$$\times\left\{1-G\left(\int_0^{Y_{ij}} e^{-\boldsymbol{\beta}_b^{\mathrm{T}}\mathbf{b}_{ij}-\boldsymbol{\beta}_w^{\mathrm{T}}\mathbf{w}_{ij}-\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{X}_{ij}(s)-R_{ij}}\mathrm{d}\Lambda(s)\right)\right\}^{1-\Delta_{ij}}$$
$$\left.\times\,\phi_i(\mathbf{R}_i;\boldsymbol{\xi})\mathrm{d}\mathbf{R}_i\right] \quad (A.3)$$

where $\Lambda\{Y_{ij}\}$ is the jump size of $\Lambda(y)$ at $y = Y_{ij}$, i.e., the value of $\Lambda(y)$ at $y = Y_{ij}$ minus its value right before $Y_{ij}$. The resulting estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}_b, \widehat{\boldsymbol{\beta}}_w, \widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\xi}}, \widehat{\Lambda})$ is the non-parametric maximum likelihood estimator of $\boldsymbol{\theta}$ [Bickel et al., 1993].

The non-parametric maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ possesses the same theoretical properties as the standard parametric maximum likelihood estimator in that $\widehat{\boldsymbol{\theta}}$ is consistent, asymptotically normal and statistically efficient. Because $\Lambda(\cdot)$ is an infinite-dimensional parameter, the proof of these results involves very advanced mathematical arguments from modern empirical process

theory and semiparametric efficient theory. The interested readers are referred to Zeng et al. [2005] for a detailed proof for this kind of problem.

It can be shown that $\widehat{\Lambda}(\cdot)$ is a step function with jumps only at $Y_{(k)}$, $k = 1,\ldots,K$, where $Y_{(k)}$ is the $k$th order statistic of the distinct uncensored trait values, and $K$ is the total number of uncensored values. Thus, we maximize (A.3) over $\boldsymbol{\beta}_b, \boldsymbol{\beta}_w, \boldsymbol{\gamma}, \boldsymbol{\xi}$, and $\Lambda\{Y_{(k)}\}$ ($k = 1,\ldots,K$) through the quasi-Newton algorithm [Press et al., 1992]. The unknown transformation function $H(y)$ in equation (2) is estimated by $\widehat{H}(y) = \log \widehat{\Lambda}(y)$. For the proportional hazards model, it is convenient to apply the EM algorithm [Dempster et al., 1977] because we can take advantage of an explicit solution for $\Lambda$ in the M-step. To ensure the positiveness of the jump sizes of $\Lambda(\cdot)$ and variance parameters $\boldsymbol{\xi}$, we use the transformed parameters $\log(\Lambda\{Y_{(k)}\})$ and $\log(\boldsymbol{\xi})$ instead of $\Lambda\{Y_{(k)}\}$ and $\boldsymbol{\xi}$ in the maximization. In the calculation of the likelihood function, we approximate the integrals over $\mathbf{R}_i$ through numerical summations such as the adaptive Gaussian quadrature approximation [Pinheiro and Bates, 1995]. We can obtain a good approximation with 10 or more quadrature points. The Laplace approximation used by Pankratz et al. [2005] is a special case of the adaptive Gaussian quadrature approximation with one quandrature point and may not be accurate enough.

For the normal error distribution, we can represent the likelihood function for $\boldsymbol{\theta}$ in a computationally more convenient way. Let $E_{ij} = \log\{\int_0^{T_{ij}} e^{-\boldsymbol{\beta}_b^T \mathbf{b}_{ij} - \boldsymbol{\beta}_w^T \mathbf{w}_{ij} - \boldsymbol{\gamma}^T \mathbf{X}_{ij}(s)}\, d\Lambda(s)\}$ and $\mathbf{E}_i = (E_{i1},\ldots,E_{in_i})^T$. Then $\mathbf{E}_i$ follows a multivariate normal distribution with mean zero and variance-covariance matrix $\mathbf{V}_i = \sigma_m^2 \Sigma_{mi} + 2\sigma_p^2 \Sigma_{pi} + \mathbf{I}_i$, where $\mathbf{I}_i$ is the $n_i$-dimensional identity matrix. Without loss of generality, assume that the first $n_{ia}$ elements of the $\Delta_{ij}$ are 0 and the remaining $n_{ib} = n_i - n_{ia}$ values are 1. In this way, we partition the data for the $i$th family into two parts: the first part consists of censored observations and the second part consists of uncensored observations. Let $a$ and $b$ index the censored and uncensored individuals, respectively. The distribution of $\mathbf{E}_{ia}$ conditional on $\mathbf{E}_{ib}$ is given by

$$\mathbf{E}_{ia} | \mathbf{E}_{ib} \sim N(\boldsymbol{\mu}'_{ia}, \mathbf{V}'_{iaa})$$

where $\boldsymbol{\mu}'_{ia} = \mathbf{V}_{iab} \mathbf{V}_{ibb}^{-1} \mathbf{E}_{ib}$ and $\mathbf{V}'_{iaa} = \mathbf{V}_{iaa} - \mathbf{V}_{iab} \times \mathbf{V}_{ibb}^{-1} \mathbf{V}_{iba}$. Thus, the likelihood function for the $i$th family can be expressed as the product of the following two terms:

$$L_{ia}(\boldsymbol{\theta}) = \int_{Y_{i1}}^{\infty} \cdots \int_{Y_{in_{ia}}}^{\infty} (2\pi)^{-n_{ia}/2} |\mathbf{V}'_{iaa}|^{-1/2} \times \exp\left\{ -\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu}'_{ia})^T \mathbf{V}'^{-1}_{iaa} (\mathbf{y} - \boldsymbol{\mu}'_{ia}) \right\} d\mathbf{y}$$

and

$$L_{ib}(\boldsymbol{\theta}) = (2\pi)^{-n_{ib}/2} |\mathbf{V}_{ibb}|^{-1/2} \exp\left( -\frac{1}{2} \mathbf{E}_{ib}^T \mathbf{V}_{ibb}^{-1} \mathbf{E}_{ib} \right) \times \prod_{j=n_{ia}+1}^{n_i} \frac{e^{\boldsymbol{\gamma}^T \mathbf{X}_{ij}(Y_{ij})} \Lambda\{Y_{ij}\}}{\int_0^{Y_{ij}} e^{\boldsymbol{\gamma}^T \mathbf{X}_{ij}(s)}\, d\Lambda(s)}.$$

The likelihood function for $n$ families is then given by

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} L_{ia}(\boldsymbol{\theta}) L_{ib}(\boldsymbol{\theta}).$$

The tail probabilities in $L_{ib}(\boldsymbol{\theta})$ can be approximated by a subroutine for computing multivariate normal probabilities given by Genz [1992]. With this alternative representation, we can reduce the order of the integral for the $i$th family from $n_i$ in (A.3) to $n_{ia}$, i.e., the number of censored observations in the $i$th family.