Genome Scans With Gene-Covariate Interaction

Jie Peng

Department of Statistics, Stanford University, Stanford, CA 94305

Hsiu-Khuern Tang

Hewlett-Packard Company, Palo Alto, CA 94304

David Siegmund

Department of Statistics, Stanford University, Stanford, CA 94305

July 27, 2005

Email: dos@stat.stanford.edu; Fax: 650-725-8977

Running Head: Gene-Covariate Interaction

Abstract

Genetic models for gene-covariate interaction are described. Methods of linkage analysis that utilize special features of these models and the corresponding score statistics are derived. Their power is compared with that of simple genome scans that ignore these special feature, and substantial gains in power are observed when the gene-covariate interaction is strong. Quantitative trait mapping in randomly ascertained sibships and affected sibpair mapping are discussed. For the latter case, a simpler statistic is proposed that has similar performance to the score statistic, but does not require the estimation of nuisance parameters. Since the nuisance parameters are not estimable solely from affected sibpair data, this statistic would be much easier to apply in practice. Similarities with linkage analysis of models for longitudinal data and multivariate phenotypes are also briefly discussed. Approximations for the p-value and power are derived under the framework of local alternatives.

Keywords: Gene mapping; Quantitative trait; Gene-environment/covariate interaction.

1 Introduction

The genetic control of a complex or quantitative trait is widely thought to involve a number of loci, which may interact with one another and/or with environmental covariates. Standard genome scanning methods usually ignore these possibilities, presumably because they involve larger, more complex models and/or because of difficulties in formulating a suitable model. A number of recent papers that explicitly consider gene-gene interactions include Cox et al. [1999], Cordell et al. [1995, 2000], and Tang and Siegmund [2002].

Gene-covariate interactions involve a much richer, and more speculative, set of possible models. Blangero and colleagues have developed components of variance methods for quantitative trait analysis, which includes the possibility of including gene-environment or more generally gene-covariate interactions (cf. Blangero and Almasy [1997], Towne et al. [1997], Almasy and Blangero [1998], and references cited therein). In these papers different covariates are modeled nonparametrically by the inclusion of appropriate variance components; but neither adjustment of the significance threshold to compensate for the increased number of parameters nor the power of the resulting statistics seems to have been studied systematically. If only additive variance components are considered in a nonparametric analysis, for each quantitative trait locus (QTL) a two valued covariate requires three variance components, and more generally a covariate that can take k different values requires $k(k+1)/2$ different variance components. Since the number of variance components grows rapidly with the number of distinct values of the covariate, this approach seems to be intrinsically limited.

Gene-covariate interactions for qualitative traits are studied by Greenwood and Bull [1999], Olson [1999], Gauderman and Siegmund [2000], and Schaid et al. [2001]. The models analysed typically invoke assumptions about how the covariates affect the probability of identity by descent. They are not derived from a penetrance model for the joint effect of genotype and covariate on the phenotype, which is the customary starting point for genetic models that do not include gene-covariate interactions. In some cases covariates have been assumed to be a property of pedigrees and are used to deal with population heterogeneity. See Schaid et al. [2003] for a review of a variety of approaches and additional references.

The primary purpose of this paper is to develop genetic models for gene-covariate interaction

and to investigate the extent to which methods utilizing such models can in principle increase the power of family based linkage analysis. We consider quantitative traits from sibship data and qualitative traits involving affected sib pairs, which we study by treating the *penetrance* of the qualitative trait as a quantitative trait. We first describe a simple model for gene-covariate interaction; then we derive the score statistic for this model and compare its performance to the "naive" statistic that ignores the interactions. While the score statistic, computed under the correct model, can be expected to have the largest possible noncentrality parameter, the model involves more degrees of freedom, hence requires a larger significance threshold to control the false positive error rate, and consequently does not always lead to greater power. For qualitative traits, it turns out that there are nuisance parameters that cannot be estimated from data on affected sib pairs alone. Hence we consider a "simplified" statistic that uses some of the information provided by the covariates, and we then use the score statistics as a device for determining how much information has been lost through this process. As in Tang and Siegmund [2001, 2002] we use the framework of local alternatives employed in large sample statistical theory. This allows us to obtain computable expressions for asymptotic noncentrality parameters and hence to compare the power of different strategies. We find that when there is gene-covariate interaction, very large sample sizes can be required to detect linkage, but use of an appropriate model can lead to large savings in the necessary sample size. This finding contrasts with our earlier studies of gene-gene interaction, where we found that properly accounting for the interaction could have a beneficial but surprisingly limited effect.

Technical results concerned with genome-wide p-values and power are discussed in an Appendix.

2 Methods

Models for quantitative traits with randomly sampled pedigrees. We begin with a model for a quantitative phenotype Y studied in randomly sampled pedigrees. The same model can be used in ascertained pedigrees (cf. Peng and Siegmund [2004]; Peng [2004]); it also applies to the penetrance of a qualitative trait and will be discussed in this context below. The model for Y is

given by

$$Y = \mu + b^T w + \sum_\tau (\alpha(\tau) + w^T \gamma(\tau)) + e. \tag{2.1}$$

The summation is over different QTL τ, which we assume are unlinked, and for simplicity act additively without gene-gene interactions. The vector w denotes covariate measurements, which take a fixed value for each individual and in random samples from the population are assumed to be independent of the other random terms in the model. For example, w may include a 0-1 variable coding for smoker/non-smoker or it may contain the measured values of continuous variables. The additive genetic effects at the trait locus τ are $\alpha(\tau)$, which is insensitive to the covariate, and the vector $\gamma(\tau)$, which interacts multiplicatively with w. By changing the values of μ and b, if necessary, we can without loss of generality assume the genetic effects have mean 0. We also assume that the residual e, which can contain unmodeled genetic effects, is uncorrelated with the the explicitly modeled terms. The model could also include terms to describe dominance effects and gene-gene interactions, if these are thought to be important.

It is often convenient and simplifies the formulas derived below to assume that w has been standardized so that population moments satisfy $\mathrm{E}w = 0$ and $\mathrm{E}ww^T = I$, the identity matrix. Although the parameters b and $\sigma^2_\gamma = \mathrm{E}[\gamma(\tau)\gamma(\tau)^T]$ are not uniquely defined unless w has been standardized, our preferred standardization is not the only possibility. When we want to distinguish between the standardized and the original covariates, we write $\tilde{w}$ to denote the original covariates.

For most of the numerical examples below w is one dimensional. For the case that w is one dimensional and two valued, the model (2.1) is equivalent to a completely non-parametric model [Towne et al. 1997].

We now summarize some basic calculations of variances and covariances. Results that are conditional on w would be the same whether w is standardized or not, but for the unconditional results we have assumed that w is standardized in order to simplify the resulting formulas. Consider an arbitrary locus on the right-hand-side of (2.1). For simplicity of notation in what follows we suppress the dependence of parameters on the trait locus τ, except when we want to emphasize this dependence. Let $\sigma^2_\alpha = \mathrm{E}(\alpha^2)$, $\sigma_{\alpha\gamma} = \mathrm{E}(\alpha\gamma)$ and $\sigma^2_\gamma = \mathrm{E}(\gamma^2)$ be the locus-specific variance components. In the general case that w is a vector, $\sigma_{\alpha\gamma}$ is also a vector and σ^2_γ is a matrix. Let

$V_{\alpha\alpha} = \sum \sigma^2_\alpha$, $V_{\alpha\gamma} = \sum \sigma_{\alpha\gamma}$ and $V_{\gamma\gamma} = \sum \sigma^2_\gamma$ be the genome-wide variance components. Observe that if w is standardized the phenotypic variance is $\sigma^2_Y = \mathrm{Var}(Y) = b^2 + V_{\alpha\alpha} + \mathrm{tr}V_{\gamma\gamma} + \sigma^2_e$. For future reference we define the overall genetic heritability H^2 and the locus specific genetic heritability h^2 by $H^2 = [V_{\alpha\alpha} + \mathrm{tr}V_{\gamma\gamma}]/\sigma^2_Y$ and $h^2 = [\sigma^2_\alpha + \mathrm{tr}\sigma^2_\gamma]/\sigma^2_Y$, respectively.

Let $\nu = \nu(t)$ denote the number of alleles shared identical by descent (IBD) at the marker locus t by two sibs with phenotypes Y_1 and Y_2 and covariates w_1 and w_2. We assume that markers are fully informative. It is straightforward to use standard software for multipoint analysis to adapt our methods for partially informative markers. We expect on the basis of simulations in related models [Peng, *et al.*, 2003] that numerical results would be very similar, but attaining the same power would require very roughly 5-25% larger sample sizes (depending on the density and informativeness of markers and whether parents are also genotyped).

For standardized or non-standardized w, by calculations along the lines of Tang and Siegmund [2001, 2002] $E(Y|w) = \mu + b^T w$, and

$$\begin{aligned}
\sigma^2_w = \mathrm{Var}(Y|w) &= V_{\alpha\alpha} + 2w^T V_{\alpha\gamma} + w^T V_{\gamma\gamma} w + \sigma^2_e, \\
\mathrm{Cov}(Y_1, Y_2 \mid w_1, w_2) &= V_{\alpha\alpha}/2 + (w_1 + w_2)^T V_{\alpha\gamma}/2 + w_1^T V_{\gamma\gamma} w_2/2 + r\sigma^2_e, \qquad (2.2) \\
\mathrm{Cov}(Y_1, Y_2 \mid w_1, w_2, \nu(\tau)) &= \mathrm{Cov}(Y_1, Y_2 \mid w_1, w_2) \\
&\quad + [\alpha_0 + (w_1 + w_2)^T \beta_0 + w_1^T \gamma_0 w_2](\nu(\tau) - 1),
\end{aligned}$$

where $\alpha_0 = \sigma^2_\alpha/2$, $\beta_0 = \sigma_{\alpha\gamma}/2$, and $\gamma_0 = \sigma^2_\gamma/2$ are defined in terms of the locus specific variance components, and r is the residual correlation between siblings. The null hypothesis is that α_0, γ_0, and β_0 all equal 0. Let $R_w = \mathrm{E}w_1 w_2^T$. By taking expectations in (2.2), we obtain expressions for the correlation $\rho = \mathrm{corr}(Y_1, Y_2)$ and conditional correlation $\rho_\nu = \mathrm{corr}(Y_1, Y_2 \mid \nu)$: $\rho = [V_{\alpha\alpha}/2 + \mathrm{tr}[(V_{\gamma\gamma}/2 + bb^T)R_w] + r\sigma^2_e]/\sigma^2_Y$ and $\rho_\nu = \rho + [\alpha_0 + \mathrm{tr}(\gamma_0 R_w)](\nu - 1)/\sigma^2_Y$.

Example. Assume that at each QTL each individual has a one dimensional (not standardized) covariate $\tilde{w}$ that is either 1, with probability p, or 0, with probability $q = 1 - p$, and that the model is given by

$$Y = \tilde{\mu} + \sum \tilde{w}\tilde{\gamma} + e. \qquad (2.3)$$

In this model of interaction, genetic effects are present in individuals having the "right" covariate. If we write this in the form (2.1) with a standardized covariate w, we have $\mu = \tilde{\mu} + \sum pE\tilde{\gamma}, w = (\tilde{w} - p)/(pq)^{1/2}, \quad b = (pq)^{1/2}E\tilde{\gamma}, \quad \gamma = (pq)^{1/2}(\tilde{\gamma} - E\tilde{\gamma}), \quad \alpha = p(\tilde{\gamma} - E\tilde{\gamma})$. In this case,

$$\alpha_0 : \beta_0 : \gamma_0 = p : (pq)^{1/2} : q. \tag{2.4}$$

For a two-valued covariate this model is equivalent to a nonparametric model that (for each QTL τ) directly assigns three different variance components to sib pairs according as (i) both have the covariate value 1, (ii) both have the covariate value 0, or (iii) they have different covariate values. For covariates taking three or more values, the suggested model has fewer variance components than a nonparametric model, which must have different variance components for each covariate value separately and for all possible pairs of values.

The parameter p has no physical interpretation when w has more than two values. We continue to use it for the numerical examples in Table I below as a simple way to specify the relative values of the parameters in (2.4). Although this restricts the parameter space to be a two dimensional subspace of the three dimensional space, additional simulations (not shown) suggest that this restriction does not put an important limitation on the insights derived from that table.

Remark. In the model of the example, each individual has an associated covariate value, and we expect that corresponding covariates of sibs will be positively correlated. An assumption of 0 correlation might be appropriate if a covariate is an indicator of sex. At the other extreme, correlation of one between sibs would make the value of the covariate a characteristic of the *sibship*, which would be appropriate to model *heterogeneity*, where we assume there are different mutations occurring at varying frequencies in different families. In this special case one could combine β_0 and γ_0 into a single parameter, so two parameters at each QTL would suffice.

Naive statistic

For later comparisons we are interested in a robust version of the simple score statistic of Tang and Siegmund [2001], which does not take the covariate into account. To provide a fair comparison for our numerical studies, we assume that for the simple score statistic, we do account for the effect of the covariate on the mean by introducing it into the model as a regressor. Failure to take this covariate effect into account would result in an approximately 10%-30% increase in the

sample size because of increased phenotypic variability (data not shown). This effectively means that we can assume the regression parameter denoted by b in equation (2.1) is zero, which we do in the numerical example that follows. In order to simplify the notation, we also assume henceforth that w is one dimensional.

The robust score statistic, is obtained as follows: (i) calculate the efficient score ℓ_α for the model (2.1) assumed to have no gene-covariate interactions, (ii) calculate the conditional variance of ℓ_α given the phenotypes, then (iii) substitute for unknown segregation parameters their maximum likelihood estimates obtained under the condition $\alpha_0 = 0,\ \beta_0 = 0,\ \gamma_0 = 0$. For the simplest case of a sample of N sib pairs, the score statistic is most easily described in terms of $D = (Y_1 - Y_2)/2^{1/2}$ and $S = (Y_1 + Y_2 - 2\mu)/2^{1/2}$. (Recall that here we have taken $b = 0$; otherwise Y_i should be replaced by $Y_i - bw_i$.) The score statistic at a putative trait locus t is (cf. Tang and Siegmund [2001])

$$Z(t) = \ell_\alpha(t)/[\sum_n C_n^2/2]^{1/2}, \tag{2.5}$$

where $\ell_\alpha(t) = \sum[\nu_n(t) - 1]C_n$ and $C_n = \hat\rho/(1-\hat\rho^2) + S_n^2/[2\hat\sigma_Y^2(1+\hat\rho)^2] - D_n^2/[2\hat\sigma_Y^2(1-\hat\rho)^2]$. Since we do not know the position of the trait locus τ, for a genome scan we use $\max_t Z(t)$, where the max extends over all marker loci t. The asymptotic noncentrality of $Z(\tau)$ is

$$(N/2)^{1/2}[\alpha_0 + \gamma_0 R_w]\sigma_Y^{-2}(1+\rho^2)/(1-\rho^2)^2 \Big/ [\mathrm{E}_0 C^2]^{1/2}.$$

The value of $\mathrm{E}_0 C^2 = \mathrm{E}_0[\mathrm{E}_0(C^2|w_1, w_2)]$ depends on the true distribution of C. For the computations reported below, we take (Y_1, Y_2) to be conditionally, given (w_1, w_2), bivariate normal with variances and covariance given by (2.2).

Score statistics

We now turn our attention to statistics that incorporate the covariate w. We consider sibships of size s and proceed as in Tang and Siegmund [2001, 2002]. We adopt the working model of variance components analysis, that within each sibship $\mathbf{Y}^T = (Y_1, Y_2, \cdots Y_s)$ is conditionally multivariate normal given the environmental variables and the pairwise identity by descent counts at a trait locus. For notational simplicity, we assume that $\mu = 0 = b$ and let $\Sigma_{w,\nu}$ denote the conditional covariance matrix, the entries of which are given by (2.2). Observe that $\Sigma_{w,\nu} =$

$\Sigma_w + \alpha_0 A_\nu + \beta_0 B_{w,\nu} + \gamma_0 \Gamma_{w,\nu}$, where the elements of Σ_w are given in the first and second lines of (2.2), A_ν has entries $\nu_{i,j} - 1$ (by convention $\nu_{i,i} - 1 = 0$), $B_{w,\nu}$ has entries $(w_i + w_j)(\nu_{i,j} - 1)$, and $\Gamma_{w,\nu}$ has entries $w_i w_j (\nu_{i,j} - 1)$.

Denoting the log likelihood function by ℓ, we find by differentiation that for the linkage parameters $\alpha_0, \beta_0, \gamma_0$ the components of the score vector at a putative trait locus t under the hypothesis that $\alpha_0 = 0,\ \beta_0 = 0,\ \gamma_0 = 0$ are

$$\begin{aligned}
\ell_\alpha(t) &= 2^{-1} \sum_n \{-\mathrm{tr}(\Sigma_w^{-1} A_\nu) + \mathrm{tr}(\Sigma_w^{-1} A_\nu \Sigma_w^{-1} \mathbf{Y}\mathbf{Y}^\mathbf{T})\}, \\
\ell_\beta(t) &= 2^{-1} \sum_n \{-\mathrm{tr}(\Sigma_w^{-1} B_{w,\nu}) + \mathrm{tr}(\Sigma_w^{-1} B_{w,\nu} \Sigma_w^{-1} \mathbf{Y}\mathbf{Y}^\mathbf{T})\}, \\
\ell_\gamma(t) &= 2^{-1} \sum_n \{-\mathrm{tr}(\Sigma_w^{-1} \Gamma_{w,\nu}) + \mathrm{tr}(\Sigma_w^{-1} \Gamma_{w,\nu} \Sigma_w^{-1} \mathbf{Y}\mathbf{Y}^\mathbf{T})\}.
\end{aligned} \tag{2.6}$$

For the segregation parameters $V_{\alpha\alpha}, V_{\alpha\gamma}$, etc., which we denote generically by s, the components of the score vector are of the form

$$\ell_s(t) = 2^{-1} \sum_n \{-\mathrm{tr}[\Sigma_w^{-1} (\partial \Sigma_w^{-1} / \partial s) + \mathrm{tr}[\Sigma_w^{-1} (\partial \Sigma_w^{-1} / \partial s) \Sigma_w^{-1} \mathbf{Y}\mathbf{Y}^\mathbf{T}]\}.$$

Since they do not involve the identity by descent counts, $\nu_{i,j} - 1$, they are uncorrelated with ℓ_α, ℓ_β, and ℓ_γ. Hence by standard likelihood theory (e.g., Cox and Hinkley [1974] p. 324), for the evaluation of noncentrality parameters in the asymptotic theory to follow, the nuisance parameters can be regarded as known.

Additional calculations yield related quantities of interest, e.g., the normalizing matrix for the robust test statistic or the Fisher information matrix F_w under the normality assumption. In general these quantities involve the entries $\sigma_w^{i,j}$ of the matrix Σ_w^{-1} or their expectations with respect to the distribution of $w^T = (w_1, \cdots, w_s)$ and cannot be exhibited explicitly, although they are easily evaluated numerically either for empirical implementation or for theoretical analysis. For example, the $\alpha\alpha$ entry of F_w is

$$I_{\alpha\alpha,w} = E(\ell_\alpha^2 | w) = 4^{-1} \sum_n \sum_{i \neq j} [(\sigma_w^{i,j})^2 + \sigma_w^{i,i} \sigma_w^{j,j}]; \tag{2.7}$$

for $I_{\alpha\gamma,w}$ each term on the right hand side of (2.7) is multiplied by $w_i w_j$; for $I_{\beta\beta,w}$ each term is multiplied by $(w_i + w_j)^2$, etc. For a robust score statistics in the spirit of Tang and Siegmund [2001], we could normalize the efficient score by conditioning (2.7) and related quantities on $\mathbf{Y}$ as well as w.

To test $\alpha_0 = \beta_0 = \gamma_0 = 0$, we first define the score vector $\ell_\theta(t) = (\ell_\alpha(t), \ell_\beta(t), \ell_\gamma(t))^T$. Let $Q(t)$ be the quadratic form derived from $\ell_\theta(t) F_w^{-1} \ell_\theta(t)$, by (i) replacing unknown variance components in F_w^{-1} by their maximum likelihood estimates under the condition $\alpha_0 = \beta_0 = \gamma_0 = 0$, and by (ii) constraining the statistic by the inequalities $\alpha_0 \geq 0$ and $\gamma_0 \geq 0$. (A third constraint, $|\beta_0| \leq (\alpha_0 \gamma_0)^{1/2}$, will be ignored for mathematical simplicity.) See Appendix A for a more complete description. For a genome scan we use $\max_t Q(t)$, where the maximum is taken over all marker loci. Under the multivariate normality assumption, large sample statistical theory (e.g., Cox and Hinkley [1974], p. 324) implies the asymptotic noncentrality of $Q(t)$ at $t = \tau$ is the Mahalanobis norm of $\theta = (\alpha_0, \beta_0, \gamma_0)^T$ with respect to the Fisher information matrix F_w, i.e., $\xi_w = (\theta^T F_w \theta)^{1/2}$. When N is large, by the law of large numbers $N^{-1} F_w$ will approximately equal its expectation, taken with respect to the distribution of w, which we denote by F. For the power calculations given below the noncentrality parameter is taken to be the unconditional value $\xi = (N \theta^T F \theta)^{1/2}$.

Remarks. (a) Several other genetic problems lead naturally to similar models. For example, a bivariate phenotype naturally involves two variance and one covariance component. The noncentrality parameters of the score statistic would satisfy constraints that are similar to the constraints arising from the model (2.1). See Wang [2003] or Turner *et al.* [2004]. A similar model could be developed to deal with longitudinal data. Equation (2.1) would represent the simplest possible model, with genotypic effects that are linear in time. See Rabinowitz and Shea [1997] for a non-genetic example, also Harville [1977] and Laird and Ware [1982] for discussions of parameter estimation. (b) For mapping QTL using pedigrees ascertained through phenotypes and/or covariates of probands, the robust score statistic has the same form as before. Ascertainment corrections should be used to obtain estimates of nuisance parameters. See, for example, Peng [2004] and Peng and Siegmund [2004]. (c) Although the analysis above can be applied to general pedigrees, we have focused on sib pairs in the numerical examples to follow. Since this leads to very large sample sizes, it is useful to recall that an effective technique for reducing sample sizes

in QTL mapping is to target large sibships/pedigrees (e.g., Tang and Siegmund [2001]). (d) The preceeding score statistics are derived from the ideal case of fully informative markers. Since the efficient score (2.6) in the QTL case and the efficient score (2.9) for the affected sibpair mapping are both linear in $\nu_{ij}-1$, when markers are partially informative, the corresponding efficient scores involve replacement of the ν_{ij} by their conditional means, $\hat{\nu}_{ij} = \mathrm{E}(\nu_{ij}|M)$, where M is the marker genotype data. The naive statistic (2.5) and the simple statistic (2.10) can be altered for partially informative markers in the same way. Standardization of these statistics involves estimation of the variance of $\hat{\nu}_{ij}$ (cf. T.Cuenco *et al.* [2003]).

Qualitative traits

Suppose now that the model (2.1) gives the penetrance of a 0-1 trait and consider a sample of N independent affected sib pairs. Viewing the penetrance of the qualitative trait as a quantitative trait value, one finds from (2.2) and calculations like those of James [1971], Risch [1990], and Tang and Siegmund [2002] that the log likelihood function is

$$\sum_n \sum_{i=0}^{2} [1_{\{\nu=i\}} \log\{E(Y_1Y_2|w_1,w_2) + (i-1)[\alpha_0 + (w_1+w_2)\beta_0 + w_1w_2\gamma_0]\}]. \tag{2.8}$$

The first component of the score vector evaluated at a putative trait locus t, with $\alpha_0 = \beta_0 = \gamma_0 = 0$ is

$$\ell_\alpha(t) = \sum_n [(\nu(t)-1)/E(Y_1Y_2|w_1,w_2)]; \tag{2.9}$$

and the other components are also easily evaluated (cf. (2.6)). The expression $E(Y_1Y_2|w_1,w_2)$, which arises from conditioning on the event that the sibs are affected, is a linear combination of $1, w_1+w_2$, and w_1w_2 (cf. display (2.2)), the coefficients of which must be estimated if we are to use (2.9) and its companions ℓ_β and ℓ_γ as the basis for a test statistic. However, these coefficients, which are population parameters, cannot be estimated from data on only affecteds. For the same reasons, we do not have unbiased estimators of the mean and variance of the covariate $\bar{w}$.

Because of these difficulties in evaluating the score statistic, we consider the three dimensional *simplified* statistic obtained by neglecting the denominator in (2.9) and in the corresponding

formulas for ℓ_β, ℓ_γ, and by reverting to non-standardized covariates. Let

$$X(t) = \begin{cases} X_1(t) = \sum_{n=1}^{N} (\nu(t) - 1) f_1(\tilde{w}_1, \tilde{w}_2) \\ X_2(t) = \sum_{n=1}^{N} (\nu(t) - 1) f_2(\tilde{w}_1, \tilde{w}_2) \\ X_3(t) = \sum_{n=1}^{N} (\nu(t) - 1) f_3(\tilde{w}_1, \tilde{w}_2), \end{cases} \tag{2.10}$$

where f_1, f_2 and f_3 are weight functions, to be specified later, that depend on the non-standardized covariate $\tilde{w}$. For each sib pair put

$$f(\tilde{w}_1, \tilde{w}_2) = (f_1(\tilde{w}_1, \tilde{w}_2), f_2(\tilde{w}_1, \tilde{w}_2), f_3(\tilde{w}_1, \tilde{w}_2))^T,\ f_0(\tilde{w}_1, \tilde{w}_2) = (1, \tilde{w}_1 + \tilde{w}_2, \tilde{w}_1 \tilde{w}_2)^T.$$

The three dimensional simplified statistic is the special case $f = f_0(\tilde{w}_1, \tilde{w}_2)$. We shall also be interested in the two dimensional simplified statistic obtained as the special case $f = (1, \tilde{w}_1 + \tilde{w}_2, 0)^T$. As indicated in (2.9), the score statistic is the special case of (2.10), where $f = f_0(w_1, w_2)/C(w_1, w_2)$, $C(w_1, w_2) = E(Y_1 Y_2 | w_1, w_2)$, and $w_i = (\tilde{w}_i - \mu_{\tilde{w}})/[\mathrm{Var}(\tilde{\mathrm{w}})]^{1/2}$.

When there is only a single, two-valued covariate, so the model (2.1) is equivalent to a completely nonparametric model, the noncentrality parameters of all true, unrestricted three-dimensional statistics of the form of equation (2.10) are the same, so in this special case the simplified statistic has the same noncentrality as the score statistic (proof omitted). This makes sense intuitively because each statistic has three degrees of freedom. We find in numerical examples, some of which are described below, that under a broad range of conditions the score statistic and the simplified statistic have comparable noncentrality parameters, so the simplified statistic is only marginally less powerful than the score statistic.

At unlinked markers the conditional covariance of $X(t)$ given covariates is

$$F_w = 2^{-1} \sum_n f(\tilde{w}_1, \tilde{w}_2) f^T(\tilde{w}_1, \tilde{w}_2). \tag{2.11}$$

Its expectation is NF, where

$$F = \frac{1}{2} \mathrm{E}(f(\tilde{w}_1, \tilde{w}_2) f^T(\tilde{w}_1, \tilde{w}_2) | Y_1 Y_2 = 1) = \frac{1}{2\mathrm{E}(Y_1 Y_2)} \mathrm{E}(f(\tilde{w}_1, \tilde{w}_2) f^T(\tilde{w}_1, \tilde{w}_2) C(w_1, w_2)). \tag{2.12}$$

Here we have introduced the abuse of notation $Y_1Y_2 = 1$ to indicate conditioning on the event that both sibs are affected, although strictly speaking Y_i is not an indicator of affected status but the penetrance, i.e., the conditional expectation of the indicator of affected status given genotype and covariates.

At a trait locus τ the conditional noncentrality (given covariates) of $X(\tau)$ defined in (2.10) equals

$$2^{-1}\sum \frac{f(\tilde{w}_1,\tilde{w}_2)f_0^T(w_1,w_2)}{C(w_1,w_2)}\theta =: B_w\theta,$$

say, where $\theta = (\alpha_0,\beta_0,\gamma_0)^T$. Its unconditional expectation is

$$\frac{N}{2\mathrm{E}(Y_1Y_2)}\mathrm{E}[f(\tilde{w}_1,\tilde{w}_2)f_0^T(w_1,w_2)]\theta =: NB\theta. \tag{2.13}$$

As a test statistic we consider the maximum over markers t of the quadratic form $||Z(t)|| = [X^T(t)F_w^{-1}X(t)]^{1/2}$, restricted by the requirements that the variance components be nonnegative, as described in Appendix A. Its conditional noncentrality parameter at the trait locus τ is $[\theta^T B_w^T F_w^{-1} B_w\theta]^{1/2}$, which by the law of large numbers $\sim [N\theta^T B^T F^{-1} B\theta]^{1/2}$ as $N\to\infty$.

For the three dimensional score statistic, F_w would be the Fisher information matrix and $B_w = F_w$. The conditional noncentrality parameter at τ would be $[\theta^T F_w\theta]^{1/2}$, and by the law of large numbers the unconditional noncentrality would be $\sim [N\theta^T F\theta]^{1/2}$. As noted above, this F_w contains unknown population variance components, which cannot be estimated on the basis of a sample consisting only of affected sib pairs and their covariates. Nevertheless, this case provides a standard of comparison for our simplified statistics, for which (2.11) shows that the corresponding F_w do have known values.

3 Results

For our numerical examples we assume an idealized human genome of 22 autosomes of 150 cM each and an intermarker spacing of 1 cM. We have performed similar calculations for a 10 cM intermarker spacing with very similar results.

The parameters $\alpha_0,\beta_0,\gamma_0$ satisfy the constraints $\alpha_0\ge 0, \gamma_0\ge 0$, and $\beta_0^2\le\alpha_0\gamma_0$. Using these constraints, we derive in Appendix A approximations for the p-value and for the power of our

statistics. The approximation to the p-value depends on the numerical value of an angle ψ_0, defined by $\cos(\psi_0) = \mathrm{corr}(\ell_\alpha, \ell_\gamma | \ell_\beta, w)$. If the covariance matrix of the statistic (in this case the conditional Fisher information matrix given the covariates) were diagonal, so the indicated partial correlation would be zero, this angle would be $\pi/2$, but in general it depends on the model and the covariates.

For the three dimensional score statistics in Tables I and II, we have used the value $\psi_0 = \pi$, which we show in Appendix A is an upper bound and hence provides a conservative approximation. We find that the threshold to maintain a 5% false positive rate is $b = 4.66$, and the required noncentrality for 90% power is $\xi = 5.58$. By way of contrast, for the one dimensional naive statistic, the corresponding values would be $b = 3.91$ and $\xi = 4.99$ [Tang and Siegmund, 2001]. Since these noncentrality parameters are proportional to the square root of the sample size, the higher threshold of the three degree of freedom statistic means that at a given trait locus if there is in fact no gene-environment interaction, there will be a loss of power, amounting to a loss of about 25% of the sample size ($(5.58/4.99)^2 - 1 = 0.25$).

If we are willing to assume that the partial correlation between ℓ_α and ℓ_γ given ℓ_β is non-negative, then $\psi_0 \le \pi/2$, which would lead to a lower threshold of about 4.5 and a noncentrality parameter of 5.4 to achieve 90% power. Simulations (data not shown) indicate that $\pi/2$ is often an approximately correct value, although it is not generally conservative.

For the three degree of freedom simplified statistic the threshold and noncentrality are slightly different: $b = 4.71$ and $\xi = 5.63$; for the two degree of freedom statistic they are $b = 4.38$ and $\xi = 5.39$. See Appendix A for details.

For a first numerical example we consider a quantitative trait that satisfies (2.3) and the constraint (2.4). Assume the overall heritability $H^2 = 0.5$ and the locus-specific heritability $h^2 = 0.25$ at τ. Table I shows the number of sib pairs required for the naive statistic, which ignores the gene-covariate interactions, and for the true score statistic for different values of p and R_w. (Recall that in the case that (w_1, w_2) is normally distributed p is a formal parameter that indicates the relation (2.4) among $\alpha_0, \beta_0, \gamma_0$.) For a trait having this overall heritability and locus specific heritability, if there were no gene-covariate interaction the required sample size would be ~ 2650 sib pairs [Tang and Siegmund, 2001]. From Table 1 we see that when R_w is large, sibs tend to have similar covariate values, and it makes little difference whether we include

the covariates in our statistic or not. When R_w is small, and especially when p is also small, the detection of linkage is intrinsically difficult and both methods require large sample sizes. The results are roughly the same for two-valued and normally distributed covariates.

We have also considered the possibility of using a two degree of freedom statistic consisting of only two of the three coordinates of the score statistic, but we have found that for the range of parameter values considered in Table 1 this simplification sometimes results in a substantial loss of power.

We now consider 0-1 traits studied via affected sib pairs. The results are qualitatively similar, although there are some significant differences. We tried a number of different models, and found results that were similar to what we obtained for the following basic model, which we first formulate without covariate effects and then modify in different ways by multiplying genetic effects of the basic model by nonstandardized covariates. For the basic model there are two unlinked disease susceptibility loci. The penetrances are additive at each locus, and each disease susceptibility allele has a population frequency of 0.05 with a penetrance of 0.25. There are no phenocopies. If this were a simple heterogeneous trait, without covariate interaction, a 1 cM genome scan based on the one dimensional score statistic would require a sample size of about 410 to have power 0.9 to detect each locus individually (hence power 0.81 to detect both loci).

For one variant, we considered a simple heterogeneity model, obtained by assuming that the covariate effectively splits the population into identifiable subgroups, with each subgroup more likely to be linked to a particular locus. In particular, we take the model (2.3), with $\tilde{w}$, the 0-1 valued covariate before standardization, multiplying the penetrance of one locus and $1-\tilde{w}$ multiplying the penetrance of the other, so (2.3) becomes $Y = \tilde{\mu} + \tilde{w}\tilde{\gamma}_1 + (1-\tilde{w})\tilde{\gamma}_2 + e$. We assume $r = 0$, so, in particular, there are no residual genetic effects. For such locus heterogeneity it is natural to assume the covariate is an attribute of the pedigree, i.e., that $R_w = 1$. In this case the true score statistic would be two dimensional, so we consider only that case. It turned out that the score statistic (essentially equivalent in this case to the two dimensional simplified statistic) is substantially better than the naive statistic over most of the range of parameters tested and never more than marginally worse. For example, suppose $p = 1/2$ and the mean and variance of $\tilde{\gamma}_1$ are the same as those of $\tilde{\gamma}_2$. Then the naive statistic requires a sample size of 292 sib pairs to achieve 90% power, while the two dimensional score statistic needs only 170.

For a second example, we assume that the first locus is affected by a two-valued covariate, as in the preceding example, but the second locus does not interact with the covariate, so (2.3) becomes $Y = \tilde{\mu} + \tilde{w}\tilde{\gamma}_1 + \tilde{\gamma}_2 + e$. We again assume that $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ have the same first two moments, and $r = 0$. In this case the naive statistic performs relatively better, although it can still be very much inferior to the score statistic to detect the covariate sensitive first locus. Results are given in Table II, where we have also included the two dimensional simplified statistic, based on just the first two coordinates of the three dimensional simplified statistic. Throughout most of the range of the table the two dimensional statistic is actually better than the three dimensional score statistic. This occurs because the noncentrality of the two dimensional statistic is almost as large as that of the score statistic, while the significance threshold is smaller. This differs from the results reported in Table I, where we were unable to find a two dimensional statistic that seemed competitive with the score statistic.

Note also that while the gene-covariate interaction, however it is treated, requires large, sometimes very large sample sizes for successful detection of Locus 1, surprisingly small sample sizes suffice to detect Locus 2, which is not affected by the covariate. Although the naive statistic always outperforms the other statistics at Locus 2, those statistics do derive a small amount of information from the covariate. If we modify the model by adding the possibility that the penetrance is affected by the environmental variable alone by including a term of the form $\tilde{b}\tilde{w}$ in the penetrance, then it is possible for the score statistic to outperform the naive statistic at Locus 2, even though the penetrance at Locus 2 is itself unaffected by the covariate (data not shown).

When the covariate is normally distributed (and we suspect for other continuous unimodal distributions as well), we found, in contrast to the results in Table I, very little gain in using any multidimensional statistics. This is presumably because a normally distributed covariate with variance sufficiently small that the penetrance remains restricted to [0,1] does not sufficiently stratify the data. For continuous bimodal and for both discrete and continuous trimodal distributions for the covariate, the results were similar to those in Table II, although the sample sizes at Locus 1 could be substantially larger. For example, for a discrete trimodal distribution concentrated at 0,1/3,1 with probabilities 0.25, 0.5, and 0.25, and $R_w = 0$, the sample sizes at Locus 1 for the naive, simplified two dimensional, simplified three dimensional and score statistics are respectively 4209, 2685, 2732, and 2607. A more complete description of these models and

detailed numerical results are given in Peng [2004].

4 Discussion

In this paper we have discussed models for gene-covariate interaction and statistical methods to exploit special features of these models. For the simplest of these models, involving a single covariate, there are three variance components (instead of one for a model with no covariate interaction), and their overall contribution to the trait variance is a quadratic function of the covariates. We have developed approximations, which may be slightly conservative, for the thresholds required to maintain a desired genome wide false positive error rate. Because of the larger number of parameters the significance thresholds must be higher, and hence a larger noncentrality is required before a significant gain in power is realized. If the components of variance associated with the interaction are negligible, there can be a loss of power equivalent to a loss of roughly 25% of the sample size relative to the naive statistic (15% if a less conservative threshold or if the two dimensional statistic is used). If the gene-covariate interaction is strong and correctly modeled, a substantial gain in power is achieved. For quantitative traits studied via random samples of sibships, some numerical exploration of the parameter space shows that one cannot in general reduce the dimensionality of the statistics without a significant loss of power in some cases.

Although our basic model in (2.1) is quadratic, by a Taylor series expansion it can be regarded as a rough approximation to a much more complex model. Because of the potential loss of power if the covariates are inappropriate or if they are measured on an inappropriate scale, and the potentially larger gain in power if the interaction is large and the model appropriate, model selection will play an important role in translating these results into practice. For quantitative traits, model fitting at a genomewide level requires only phenotype and covariate data, and can be based on standard statistical software. See, for example, Rabinowitz and Shea [1997], who, adapting methods of Harville [1977] and Laird and Ware [1982], apply our segregation model in a (non-genetic) longitudinal setting. The problem of covariate selection here is similar to the same problem in multiple regression analysis, and many of the same ideas can be applied.

For qualitative traits using affected sib pairs, we have observed that certain nuisance parameters entering into the likelihood function cannot be estimated on the basis of data from affected sib

pairs alone, which complicates both model selection and linkage analysis. To deal with the latter problem we have suggested two and three dimensional simplified statistics that do not involve estimation of those parameters. These simplified statistics are related to what Schaid *et al.* [2003] call regression statistics, although we do not make any assumption of homoscedasticity. The two dimensional simplified statistic is similar to the linear trend statistic favored by Gauderman and Siegmund [2000]. For the range of models we have explored numerically, we find that the three dimensional simplified statistic is usually about as powerful as the score statistic, and the two dimensional simplified statistic can be more powerful than the three dimensional simplified statistic. This latter conclusion appears to be roughly consistent with the conclusions of Gauderman and Siegmund [2000].

In comparison with gene-gene interaction, the problem of gene-covariate interaction is substantially more difficult, since the number of covariates and the number of ways they can interact is essentially unlimited. However, the possibilities to gain substantial statistical power by correctly modeling gene-covariate interaction appear to be greater. To obtain some insight into this difference, consider a quantitative trait and suppose there is additive-additive gene-gene interaction between two unlinked genes. According to Tang and Siegmund [2002], the noncentrality parameter of the one dimensional score statistic to test for a marginal effect at a single locus has a noncentrality parameter at the trait locus proportional to $\sigma_\alpha^2/2 + \sigma_{\alpha\alpha}^2/4$, where σ_α^2 denotes the locus specific additive variance and $\sigma_{\alpha\alpha}^2$ denotes the additive-additive interaction variance. This can be compared with the noncentrality parameter of the naive statistic of Section 2, which is proportional to $\sigma_\alpha^2/2 + R_w\sigma_\gamma^2/2$. While the second variance component in both these expressions involves the interaction and can be larger than the additive variance component of the first term, in the case of gene-gene interaction, the coefficient of the interaction variance component comes from correlation of identity-by-descent at the two loci, which is determined by Mendelian segregation. A certain fraction of the interaction variance component forms part of the noncentrality of the marginal statistic, and the amount of additional linkage information that can be obtained by explicitly modeling the gene-gene interaction is limited. The correlation R_w is not restricted, and as our numerical examples show, especially when R_w is small, the impact of the interaction variance component on the noncentrality parameter of the naive statistic is also small, and there is substantial additional linkage information to be extracted by using a multidimensional statistic.

Finally, we note that our results indicate yet another possible reason for the acknowledged difficulty in replicating linkage findings for complex diseases. If there is gene-covariate interaction, covariate heterogeneity between study populations can have a large effect on noncentrality parameters, so that a locus that is comparatively easily detected in one population could be difficult to detect in another.

Acknowledgments This research was supported by NIH Grant RO1 HG00848 and by a Stanford Graduate Fellowship.

Appendix A: Significance level and power

In this appendix, we state some approximations for the p-value and power of our scan statistics. Similar results for one and two degrees of freedom are obtained by Feingold et al. [1993] and Dupuis and Siegmund [2000]. Additional details are given in Tang [2000].

It will be convenient to write the parameters in the order $\theta = (\beta_0, \alpha_0, \gamma_0)^T$. By large sample theory, at a true trait locus τ, the score vector $\ell_\theta(\tau) := (\ell_\beta(\tau), \ell_\alpha(\tau), \ell_\gamma(\tau))^T$ is approximately normal with mean vector $F\theta$ and covariance matrix F, where $F = (I_{ij})$ denotes the Fisher information matrix. (Here and below, for notational convenience we are suppressing dependence on the covariates.) Let $F = LL^T$ be the Cholesky decomposition of F. Then $\mathbf{Z} = (Z_1, Z_2, Z_3)^T = L^{-1}\ell_\theta$ is approximately normal with mean vector $\boldsymbol{\xi} = L^T\theta$ and identity covariance matrix. As processes, $Z_i(t)$, $i = 1, 2, 3$, are (asymptotically) independent Gaussian processes.

To find the lower triangular matrix L, write $F = DRD$, where R is the correlation matrix of ℓ_θ, with entries $r_{\alpha\beta} = \mathrm{corr}(\ell_\alpha, \ell_\beta)$, *etc.*, and $D = \mathrm{diag}(I_{\beta\beta}^{1/2}, I_{\alpha\alpha}^{1/2}, I_{\gamma\gamma}^{1/2})$. By forming the Cholesky decomposition of R, we get

$$L^T = \begin{bmatrix} 1 & r_{\alpha\beta} & r_{\beta\gamma} \\ 0 & (1 - r_{\alpha\beta}^2)^{1/2} & (r_{\alpha\gamma} - r_{\alpha\beta}r_{\beta\gamma})/(1 - r_{\alpha\beta}^2)^{1/2} \\ 0 & 0 & [1 - r_{\beta\gamma}^2 - (r_{\alpha\gamma} - r_{\alpha\beta}r_{\beta\gamma})^2/(1 - r_{\alpha\beta}^2)]^{1/2} \end{bmatrix} D.$$

From this, we see that the constraints $\alpha_0 \geq 0$, $\gamma_0 \geq 0$ translate into $\xi_3 \geq 0, \quad \xi_2 \geq \xi_3 \cot\psi_0$, where

$\psi_0 \in [0, \pi]$ satisfies

$$\cos\psi_0 = \frac{r_{\alpha\gamma} - r_{\alpha\beta}r_{\beta\gamma}}{(1 - r_{\alpha\beta}^2)^{1/2}(1 - r_{\beta\gamma}^2)^{1/2}} = \mathrm{corr}(\ell_\alpha, \ell_\gamma \mid \ell_\beta). \tag{A.1}$$

These constraints define a region in $\mathbf{Z}$-space whose projection on the unit sphere is contained in a region A bounded by two semi-circular arcs of great circles. These arcs may be thought of as longitudinal lines on the Earth separated by ψ_0 radians. There is an additional constraint that follows from the inequality $\beta_0^2 \leq \alpha_0\gamma_0$, which reduces the size of the true parameter space to a subset $\tilde{A}$ of A. We will neglect this third constraint whenever $\arg \mathbf{Z} \in A$. For a genome wide search, we use the constrained statistic $\max_t Q(t)$, where $Q = \|\mathbf{Z}\|$ for $\arg \mathbf{Z} \in A$, and is otherwise the length of the one dimensional projection of $\mathbf{Z}$ onto the cone defined by $\tilde{A}$. Since the region A is larger than strictly necessary, this leads to a conservative approximation.

Assume that markers are equally spaced at intermarker distance Δ cM. Using the method developed by Feingold et al. [1993] and Dupuis and Siegmund [2000], we obtain for a single chromosome of length ℓ cM the approximation

$$\mathrm{P}\{\max_i Q(i\Delta) > b\} \approx \ell b^2 e^{-b^2/2}\beta\nu\{(2/\pi)^{1/2}(\psi_0/2\pi)b + 1/[b(2\pi)^{1/2}]\},$$

where $\beta = 0.04/\mathrm{cM}$ and $\nu = \nu[b(2\beta\Delta)^{1/2}]$. The function ν is a special function that can be easily computed numerically. For relatively small x, say $0 < x < 2$, it is given approximately by $\nu(x) \approx \exp(-0.583x)$. See Siegmund [1985]. One can improve this approximation slightly by adding as an end correction for the first marker on a chromosome the amount $(\psi_0/2\pi)\mathrm{P}(\chi_3 > b) + (1/2)\mathrm{P}(\chi_1 > b)$, where χ_k denotes a χ random variable with k degrees of freedom. If we denote by Q_c the single chromosome approximation, the whole genome approximate p-value would be $1 - \exp(-\sum_c Q_c)$.

For the numerical results of Tables I and II we have taken $\Delta = 1$ cM and have used the conservative value $\psi_0 = \pi$, leading to the threshold of 4.66 for the three dimensional score statistics. Using the approximation for the power given below, we find that a noncentrality parameter of about 5.58 is required to guarantee 90% power. If we are willing to assume that the partial correlation in (A.1) is non-negative, then $\psi_0 \leq \pi/2$, which would lead to a threshold of about

4.5 and a noncentrality parameter of 5.4. In practice one might estimate ψ_0 (as a function of the covariate values).

For dealing with bivariate traits linked to the same locus, as outlined in the Methods section, we have found numerically that $\psi_0/2$ is usually close to $\pi/2$. This is consistent with the numerical calculations of Wang [2003]. Turner et al. [2004] have stated that $\pi/2$ is the appropriate value. Although this appears not to be mathematically correct, it does seem adequate in practice.

A slightly different argument leading to weaker constraints is required for the three dimensional simplified statistic defined in (2.10). Let $\theta = (\alpha_0, \beta_0, \gamma_0)^T$. The asymptotic noncentrality parameter of $\mathbf{Z}(\tau) = L_w^{-1} X(\tau)$ is proportional to $L^{-1}B\theta$, where F defined in (2.12) has LL^T as its Cholesky decomposition and B is the matrix defined in (2.13). The first row of B is proportional to $(1, 0, R_w)$ with a positive constant of proportionality. If we assume, as seems reasonable that $R_w \geq 0$, then the first entry in $B\theta$ is $\alpha_0 + R_w\gamma_0 \geq 0$, and hence the first entry in $L^{-1}B\theta$ is nonnegative. This leads to the constraint on $\mathbf{Z}$ that its first entry be nonnegative, which amounts to putting $\psi_0 = \pi$ in the preceding case. When $Z_1 < 0$, we suggest projecting $\mathbf{Z}$ onto the two dimensional boundary (although when $R_w > 0$, one might consider a projection to the origin). A similar argument applies to the two dimensional simplified statistic. The genome wide significance thresholds for the three and two dimensional simplified statistics are 4.71 and 4.38, which require noncentrality parameters of 5.63 and 5.39, respectively, to achieve 90% power.

We calculate the power under the alternative $\boldsymbol{\xi} = \mathrm{E}\mathbf{Z}(\tau) \neq \mathbf{0}$. The resulting formula is quite simple in the case that the QTL is at a marker. For simplicity, we also assume that $\arg \boldsymbol{\xi}$ lies in the interior of A, which will be the case when our model is correct, so α_0, β_0 and γ_0 are all different from 0. See Dupuis and Siegmund [2000] for the kind of modification required when this is not the case. From the decomposition

$$\mathrm{P}\{\max Q(i\Delta) > b\} = \mathrm{P}\{Q(\tau) > b)\} + \mathrm{E}[\mathrm{P}\{\max Q(i\Delta) > b \mid \mathbf{Z}(\tau)\}; Q(\tau) \leq b], \tag{A.2}$$

we obtain as an approximation for the power

$$1 - \Phi(b - \xi) + \frac{1}{\xi}\varphi(b - \xi) + \frac{b}{\xi}\varphi(b - \xi)\Big[\frac{2\nu}{\xi} - \frac{\nu^2}{b + \xi}\Big],$$

where $\nu = \nu[b(2\beta\Delta)^{1/2}]$.

REFERENCES

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

Blangero J, Almasy L (1997). Multipoint oligogenic linkage analysis of quantitative traits. Genet. Epidemiol. 14: 959-964.

Cordell HJ, Todd JA, Bennett ST, Kawaguchi Y, Farrall M (1995). Two-locus maximum Lod score analysis of a multifactorial trait: joint consideration of IDDM2 and IDDM4 with IDDM1 in Type 1 diabetes. Am. J. Hum. Genet. 57: 920-34.

Cordell, HJ, Wedig, GC, Jacobs, KB and Elston, R (2000). Multilocus linkage tests based on affected relative pairs, Am. J. Hum. Genet. 66: 1273-1286.

Cox DR, Hinkley DV (1974). Theoretical Statistics, Chapman and Hall, London.

Cox NJ, Frigge M, Nicolae DL, Concannon P, Hanis CL, Bell GI, Kong A (1999). Loci on chromosome 2 (NIDDM1) and 15 interact to increase susceptibility to diabetes in Mexican Americans. Nat. Genet. 21: 213-215.

Dupuis J, Siegmund D (2000). Boundary crossing probabilities in linkage analysis. Game Theory, Optimal Stopping, Probability and Statistics: Papers in honor of Thomas Ferguson, F. Thomas Bruss and L. LeCam, eds., Institute of Mathematical Statistics, Beechwood, OH.

Feingold, E., Brown, P.O., and Siegmund, D. (1993). Gaussian models for genetic linkage analysis using complete high resolution maps of identity-by-decent. Am. J. Hum. Genet. 53: 234-251.

Gauderman, W, Siegmund, K (2000). Gene-environment interaction and affected sib pair linkage analysis. Hum. Hered. 52: 34-46.

Greenwood CMT, Bull, S B (1999). Analysis of affected sib pairs with covariates–with and without constraints. Am. J. Hum. Genet. 64: 871-885.

Harville DA (1977). Maximum likelihood approaches to variance component estimation and to related problems, Jour. Amer. Statist. Assoc. 72: 384-395.

James JW (1971). Frequency in relatives for an all-or-none trait, Ann. Hum. Genet. 35: 47-48.

Kruglyak L, Lander ES (1995). Complete multipoint sib pair analysis of qualitative and quantitative traits. Am. J. Hum. Genet. 57: 439-454.

Laird, N and Ware, JH (1982). Random-effects models for longitudinal data, Biometrics 38: 963-974.

Olson JM (1999). A general conditional-logistic model for affected-relative-pair linkage analysis. Am. J. Hum. Genet. 65: 1760-1769.

Peng J (2004). Score Statistics to Map Genes in Humans. PhD thesis, Stanford University, USA.

Peng J and Siegmund D (2004). Mapping quantitative traits with random and with ascertained sibships, Proc. Natl. Acad. Sci USA 101: 7845-7850.

Peng J, Tang H-K, Siegmund, D (2003). Information content of multi-point analysis and efficiency of two stage genotyping, Am. J. Hum. Genet., Suppl. 73: 608.

Rabinowitz D, Shea S (1997). Random effects analysis of children's blood pressure data. Statist Sci 12: 185–194.

Risch, N (1990). Linkage strategies for genetically complex traits I, The power of affected relative pairs, Am. J. Hum. Genet. 46: 222-228.

Schaid DJ, McDonnell SK, Thibodeau SN (2001). Regression models for linkage heterogeneity applied to familial prostate cancer. Am. J. Hum. Genet. 68: 1189-1196.

Schaid DJ, Olson JM, Gauderman WJ, Elston RC (2003). Regression models for linkage: Issues of Traits, covariates, heterogeneity, and interaction. Hum. Hered. 55: 86-96.

Siegmund D (1985). Sequential Analysis: Tests and Confidence Intervals, Springer-Verlag, New York.

Tang H-K (2000). Using variance components to map quantitative trait loci in humans. PhD thesis, Stanford University, USA.

Tang H-K and Siegmund D (2001). Mapping quantitative trait loci in oligogenic models. Biostatistics 2: 147-162.

Tang H-K Siegmund D (2002). Mapping multiple genes for complex or quantitative traits. Genet. Epidemiol. 22: 313-327.

T.Cuenco, K., Szatkiewicz, J.P. & Feingold, E. (2003) Recent advances in human quantitative-trait-locus mapping: comparison of methods for selected sibling pairs. Am. J. Hum. Genet. 73, 863-873.

Towne B, Siervogel RM, Blangero J (1997). Effects of genotype-by-sex interaction on quantitative trait linkage analysis. Genet. Epidemiol. 14: 1053-1058.

Turner ST, Kardia SLR, Boerwinkle E, de Andrade M (2004). Multivariate linkage analysis of blood pressure and body mass index. Genet. Epidemiol. 27: 64-73.

Wang K (2003). Mapping quantitative trait loci using multiple phenotypes in general pedigrees. Hum. Hered. 55: 1-15.

Table I: Quantitative Traits: Number of sib pairs for naive and score statistics.

The general model is given by equation (2.1). We assume that $H^2 = 1/2$ and $h^2 = 1/4$ at the primary locus. The column headed "naive" is the sample size for the statistic (2.5). The columns headed "two-point" are the sample sizes when the covariate is two-valued. The columns headed "normal" are the sample sizes when the standardized covariates are bivariate normal within sib pairs with correlation R_w, and the noncentrality parameters satisfy relation (2.4). For this case the noncentrality parameters are determined by simulations with 10^5 samples.

			Two-point		Normal	
p	R_w	ρ	Naive	Score	Naive	Score
0.75	1.0	0.25	2792	2960	3044	3460
	0.9	0.244	2953	3254	3180	3576
	0.5	0.219	3728	3626	3849	4083
	0.1	0.194	4801	4093	4798	4690
	0	0.188	5134	4230	5096	4866
0.5	1.0	0.25	3104	2975	3332	3604
	0.9	0.238	3473	3356	3658	3824
	0.5	0.188	5682	4251	5725	4717
	0.1	0.138	10430	5797	10507	5680
	0	0.125	12508	6377	12677	5942
0.25	1.0	0.25	4038	3534	3504	3698
	0.9	0.231	4781	4095	4056	4030
	0.5	0.156	10388	6061	8782	5267
	0.1	0.081	33761	11655	32729	6268
	0	0.063	53741	15152	55376	6456

Table II: Qualitative Traits: One Locus Interacts.

Locus 1 interacts with a two-valued covariate. The penetrance of the second locus is unaffected by the covariate. For other details of the model see the text. The entries under the columns "Naive", "2d-simp" and "3d-score" are sample sizes required for 90% power in a genome scan with markers spaced at 1 cM for the one dimensional naive, two dimensional simplified, and three dimensional score statistics, as defined in the text. Note that in this case, since there is only a single two-valued covariate, the three dimensional score statistic and the three dimensional simplified statistic have essentially the same power.

			Locus 1			Locus 2		
p	R_w	ρ	Naive	2d-simp	3d-score	Naive	2d-simp	3d-score
0.75	1.0	0.255	525	536	536	295	301	301
	0.9	0.252	542	548	584	290	296	316
	0.5	0.239	623	599	626	268	276	294
	0.1	0.226	733	654	679	248	256	273
	0	0.223	767	668	694	243	251	268
0.5	1.0	0.259	818	669	669	204	200	200
	0.9	0.254	879	703	744	198	195	209
	0.5	0.234	1245	883	885	175	177	189
	0.1	0.214	2026	1199	1129	153	160	171
	0	0.208	2367	1318	1220	148	156	166
0.25	1.0	0.259	2087	1069	1069	130	133	133
	0.9	0.254	2371	1162	1221	127	130	140
	0.5	0.236	4260	1811	1705	113	120	129
	0.1	0.2181	15086	4149	3080	100	111	118
	0	0.213	24683	5983	3940	96	108	115