17

MULTIVARIATE QTL ANALYSIS USING STRUCTURAL EQUATION MODELLING: A LOOK AT POWER UNDER SIMPLE CONDITIONS

Dorret I. Boomsma
Conor V. Dolan

ABSTRACT

In linkage analysis of quantitative, complex, traits the power to detect loci that explain a small to medium proportion of the genetic variance is problematic. In this paper we address the question how genetic analysis of multivariate data can be employed to increase the power to detect a quantitative trait locus using identity-by-descent mapping in sibling pairs, or dizygotic twins. These analyses are carried out with structural equation modeling, using the Mx computer program. Power calculations show that structural equation modeling is superior to the Haseman and Elston regression method. Furthermore, the power to detect a QTL can be substantially increased by considering multiple indicators of the phenotypic trait of interest. In the models used the gain in power beyond three or four indicators was, however, minimal. Detection of a dominant gene effect was shown to be unrealistic because of the large numbers needed.

Structural equation modeling or genetic covariance structure modeling (GCSM), provides a general and flexible approach to analyze data gathered in genetically informative samples[1,2]. In applying GCSM to such data, genotypic and environmental effects are modeled as the contribution of latent (unmeasured) variables to the (possibly multivariate) phenotypic individual differences. These latent factors represent the effects of many unidentified influences. In the case of a genetic factor, these effects are due to a possibly large, but unknown, number of genes (polygenes). The contributions of the latent variables are estimated as regression coefficients in the linear regression of the observed variables on the latent variables. Given an appropriate design, providing sufficient information to identify these regression coefficients, actual estimates may be obtained using a number of well disseminated computer programs, such as LISREL[3], or Mx[4] (Neale, 1997). These programs allow estimation of parameters by means of a number of estimators including normal theory maximum likelihood (ML) and weighted least squares (WLS). The latter can be applied to analyze correlations among discrete variables (e.g. tetrachoric or polychoric correlations) and nonnormal variables. A very useful estimator in the Mx program is the normal theory raw data likelihood estimator. This estimator enables one to handle missing data and to model selected samples.

Identification of quantitative genetic models is achieved, for example, by inclusion of monozygotic (MZ) and dizygotic (DZ) twins into the study. MZ twins are genetically identical while DZ twins (and siblings) share on average 50% of their segregating genes. If MZ twins are found to resemble each other more closely than DZ twins, this suggests that genetic influences are contributing to the phenotypic individual differences in the trait under consideration. One advantage of GCSM is that this approach can be generalized readily to multivariate and longitudinal data. Just as twin data can be used to decompose the variance for a single trait into a genetic and a non-genetic part, multivariate twin data can be used to decompose the covariance between traits, or between repeated measures of the same trait, into a part due to genetic covariance and a part due to environmental covariance between variables[1,5].

DETECTION OF QTLS

The flexibility of GCSM is also evident in the relative ease with which observed genotypic or environmental information can be incorporated into the analysis. An important recent development involves the incorporation of genetic information derived from marker data, which makes it possible to detect quantitative trait loci[6-8]. A quantitative trait locus (QTL) represents a stretch of a chromosome, which includes a segregating gene that contributes to individual differences in the phenotype of interest. The segregating gene has a relatively large contribution to the phenotypic variance compared to the contributions of each polygene making up a genetic latent variable. However, compared to the total effects of the polygenetic and environmental effects, the effect of the QTL may be quite small. For instance, the QTL may account for a mere 5%, or 10% of the phenotypic variance. In GCSM, the QTL is treated in the same way as a polygenetic or an environmental factor, i.e., as a latent variable. The relationship between the QTL and the phenotypic individual differences is also modeled using linear regression. The correlation between QTL factors of siblings is obtained from measured genotypic (marker) data.

The simultaneous analysis of DNA marker data and phenotypic information from sib-pairs, or dizygotic twins, to test for the presence of a QTL was developed by Haseman and Elston[9]. In addition to the measured phenotype in the sib-pairs, the Haseman and Elston method requires data relating to the siblings' genotypes at specific loci in the vicinity of the QTL. Such loci serve as markers, i.e., genetic polymorphisms with known and detectable alleles. Using the marker data, it is possible to establish the expected proportion of alleles at a given marker locus that the sibs share identical by descent (IBD, see below). The Haseman and Elston method[9] involves regressing the squared phenotypic difference score of the sibs on this proportion.

The detection of a QTL has been viewed as problematic, because of its expected relatively small effect size, and the requirement of extensive (and expensive) marker data. However, several developments have made QTL analysis feasible: the availability of marker sets consisting of many highly informative markers distributed throughout the genome (and the increasingly cheap methods of marker typing); the development of multi-point mapping methods to obtain optimal estimates of IBD status throughout the genome[10-12]; the development of selective sampling strategies to identify the most informative sib-pairs[13-18]; and, finally, the replacement of the Haseman and Elston regression method with genetic covariance structure modeling[6-8].

The use of GCSM, instead of the Haseman and Elston regression method, allows one to model the effects of a single QTL on the bivariate distribution of the sib-pairs, and to simultaneously analyze multiple indicators of a given phenotype[6]. Analyzing the bivariate distribution instead of the squared phenotypic sib-pair difference score has been shown to be more powerful[7]. As the use of multiple indicators is known to increase power in factor analysis to detect a latent factor[8], it is likely that the multiple indicators will also increase the power to detect the presence of a QTL.

MODELS

In this chapter we investigate how the use of multivariate data, compared to univariate data, increases the statistical power to detect a QTL. Multivariate data can be collected by measuring the same variable at different time points or by measuring different (correlated) variables at the same time point. The present power calculations supplement those presented in Boomsma and Dolan[20]. Boomsma and Dolan [20] considered 3 and 4 indicator models and two linear combinations of the indicators. Their calculations are limited to a codominant QTL. Here we also consider a 4 indicator model. However, we consider a dominant QTL in addition to a codominant QTL, and we investigate the effects on the power of introducing additional indicators to the model. The specific design that we focus on in this chapter is one in which the same trait is measured repeatedly across time. We assume that the time-interval between measurement occasions is short and that observed phenotypic individual differences are due to the same genes (QTL and background genetic effects) at each time-points, and that no new genetic influences are expressed across time. Measurement error (or time-specific environmental influences) thus is the only source of discontinuity across time.

Before introducing the models employed in the power calculations, we explain briefly the meaning of the term 'identity by descent' (IBD), as this is a central concept in QTL analysis. The two parents of a sib-pair are characterized by two alleles at each marker locus (say, A_i, A_j and A_k, A_l). Each member of a sib-pair

inherits a single allele from his mother (A_i or A_j) and a single allele from his father (A_k or A_l). The sib-pairs may both have inherited the same allele from their father and the same allele from their mother (e.g. A_iA_k and A_iA_k). In this case, the sib-pair is characterized by IBD status 2 at the marker locus. Alternatively, the sibs may share the same allele from the mother (Ai), but each sibling inherited a different allele from the father (A_k and A_l; resulting in genotypes A_iA_k and A_iA_l in the offspring). They are then characterized by IBD status 1. Finally, they may have inherited a different allele from the father and a different allele from the mother. In this case they are IBD 0 at the marker locus (A_iA_k and A_jA_l). The reader is referred to Table 17.2 in Haseman and Elston[9] for an exhaustive list of possibilities. Note that:

1. IBD status is a characteristic of a sib-pair, not of an individual sibling;
2. IBD status at a given marker may be hard, if not impossible, to establish if, for example, alleles of the parents are identical (see Haseman and Elston[9] Table 17.2);
3. a parent and child have by definition IBD status 1 and MZ twins have IBD status 2 across all loci;
4. IBD status tells you nothing about the actual genotype of the sib-pairs.

If a marker is situated at a large distance from the QTL, the IBD status at the marker locus will be uninformative of the IBD status at the QTL due to recombination. However, if the marker is close to the QTL, the IBD status at the marker locus can serve as a proxy for the IBD status at the QTL. The IBD status at the marker locus can then be used to determine the degree of genetic relatedness at the QTL, just as the degree of genetic relatedness between additive polygenetic values of sib-pairs is expressed by the correlation of 0.5. It is this information that is exploited in both the Haseman and Elston regression method and in structural equation modeling methods to identify the regression coefficients in the regression of the phenotype(s) on QTL.

In practice, the marker data of the sibs and, if available, from their parents, are used to estimate the proportion of alleles shared IBD by the sibs (e.g. Kruglyak and Lander, 1995). These proportions corresponding to IBD=0, IBD=1 and IBD=2, are 0, 0.5 and 1 respectively. The probability that a sibpair shares a specific proportion of alleles IBD (either p[0], p[½], or p[1]) is calculated for each sibpair conditional on their marker data. The unconditional values of these probabilities, i.e. the expected values in the population, equal p[0]=.25, p[½]=.5, and p[1]=.25.

IBD marker probabilities provide information about the contribution of the QTL to the phenotypic resemblance of the sib-pair. If the QTL is codominant, the correlation between the QTL effects of sibpair i is equal to the estimate of the mean proportion of alleles shared IBD in sibpair i, π_i, and can be given by:[21]

$$\pi_i = p[½]_i \star .5 + p[1]_i$$

The effects of a dominant QTL are modeled in two parts: an additive part and a dominant part. The additive part is represented by the so-called breeding value and the dominant part, by the dominance deviation. The correlation between the sibs in breeding value still equals π_i, but the correlation of the dominance deviations equals $p[1]_i$. In summary, the correlation between the latent variables of the sibpair i are:

latent variable	correlation
polygenic additive latent factor	0.5
unshared environmental factor	0
additive QTL part (breeding values)	π_i
QTL dominance deviation	$p[1]_i$

In practice, both π_i and $p[1]_i$ may vary between 0 and 1 (although values of π_i do constrain values of $p[1]_i$, and vice versa). As explained below, we introduce simplifying assumptions, that result in π_i and $p[1]_i$ assuming a limited number of values. This greatly facilitates the power calculations. To indicate the expected (population) value of $p[1]_i$ and π_i, we drop the subscript i. These values are π=.5 and p[1]=.25.

Haseman and Elston Model

The original Haseman and Elston[9] sib-pair approach to linkage analysis with quantitative traits estimates the regression of the squared difference between trait values of siblings on the proportion of alleles shared IBD at a marker locus:

$$Y_i = \alpha + \beta\pi_i.$$

Let P(i,j) denote the zero mean phenotypic scores of sib j (j=1,2) in sibship i (i=1,N), then Y_i equals $[P(i,1) - P(i,2)]^2$, and π_i is the proportion of alleles shared IBD by the sibs in sibship j at the marker locus. If the regression is negative and significant, this is evidence for linkage. If there is no recombination between the marker and the QTL locus, β is a direct estimate of -2Vq, where Vq is the variance attributable to the QTL. The expectation for the squared difference score of two siblings, E[Y], may be written as:

$$\begin{aligned} E[Y] &= Var(P(i,1)) + Var(P(i,2)) - 2cov(P(i,1), P(i,2)) \\ &= 2(Ve + Va + Vq) - 2(0.5Va + \pi Vq) \\ &= 2Ve + Va + 2Vq - 2\pi Vq, \end{aligned} \quad (1)$$

where Ve denotes variance due to environmental effects not shared by family members, Va denotes variance due to background genetic effects, Vq denotes the QTL variance. It may be seen from this expression that, when working with squared difference scores, Ve and Va are not separately identified.

If we consider the possibility that the effect of the QTL on the phenotype consists of an additive (codominant) genetic component and a non-additive (recessive or dominance) part, the expectation for Y can be written as:

$$E[Y] = 2Ve + Va + 2Vq + 2Vd - 2\pi Vq - 2p[1]Vd, \quad (2)$$

where Ve again denotes environmental variance, Va, the variance due to background genetic effects, and Vq and Vd now represent additive and non-additive genetic variance attributable to the QTL. Equation (1) is usually fitted by ordinary least squares, and the significance of the parameter β is established by means of the t-test.

GENETIC COVARIANCE STRUCTURE MODELING INCLUDING A QTL

A structural equation modeling approach to QTL analysis with univariate sib-pair data involves the model:

$$P(i,j) = \lambda_a A(i,j) + \lambda_e E(i,j) + \lambda_q Q(i,j) + \lambda_d D(i,j) \quad (3)$$

where P(i,j), is a function of the sibs additive QTL value (Q), non-additive QTL value (D), the scores on the latent genetic background (A) and on the environmental factor (E). A path diagram for this model is given in Figure 17.1. In this model, we assume that all variables have zero mean. We also assume that the latent variables (A, E, Q, D) are standardized, so that the phenotypic variances and covariances only depend on the regression coefficients (λ_a, λ_e, λ_q, λ_d). Finally, we assume that the latent variables are uncorrelated. As shown in equations 4 and 5, these parameters express the influence of the latent variables on the phenotype.

$$Var[P(i,1)] = Var[P(i,2)] = \lambda_a^2 + \lambda_e^2 + \lambda_q^2 + \lambda_d^2 \quad (4)$$

$$Cov[P(i,1),P(i,2)] = 0.5\,\lambda_a^2 + \pi\,\lambda_q^2 + p[1]\,\lambda_d^2 \quad (5)$$

This model is usually fitted using a program for covariance structure modeling, such as LISREL[3] or Mx[4]. If the phenotypes are approximately normally distributed, maximum likelihood estimation can be used and the significance of the regression coefficients can be tested by means of the loglikelihood ratio test.

The structural equation modeling approach to linkage analysis of multivariate phenotypes is a generalization of the univariate case:

$$\mathbf{P}(i,j) = \mathbf{\Lambda}_a \mathbf{A}(i,j) + \mathbf{\Lambda}_e \mathbf{E}(i,j) + \mathbf{\Lambda}_q \mathbf{Q}(i,j) + \mathbf{\Lambda}_d D(i,j) \quad (6)$$

Here $\mathbf{P}(i,j)$ represents the (px1) random vector of phenotypic (centered) scores of sib j in sibship i. The ($p \times n_a$) matrix $\mathbf{\Lambda}_a$ contains regression coefficients relating the p phenotypes to n_a latent additive genetic factors in the $n_a \times 1$ vector $\mathbf{A}(i,j)$. The matrices $\mathbf{\Lambda}_e$, $\mathbf{\Lambda}_q$, and $\mathbf{\Lambda}_d$ are defined in the same manner. Similarly, the vectors $\mathbf{E}(i,j)$ ($n_e \times 1$), $\mathbf{Q}(i,j)$ ($n_q \times 1$), and $\mathbf{D}(i,j)$ ($n_q \times 1$), are vectors containing unshared environmental deviation scores, QTL additive deviation scores, and QTL dominance deviation scores. As above all the deviation scores have zero mean and are standardized. The partitioned (2px2p) covariance matrix, $\mathbf{\Sigma}_i$ of the multivariate phenotypic scores $\mathbf{P}(i,1)$ and $\mathbf{P}(i,2)$ equals:

$$\mathbf{\Sigma}_i = \begin{bmatrix} \mathbf{\Sigma}_{11i} & \mathbf{\Sigma}_{2\,1i}^t \\ \mathbf{\Sigma}_{21i} & \mathbf{\Sigma}_{22i} \end{bmatrix}$$

where $\mathbf{\Sigma}_{11i} = \mathbf{\Sigma}_{22i}$. Assuming the latent variables (**A, D, E, Q**) are uncorrelated, the (pxp) covariance matrix $\mathbf{\Sigma}_{11i}$ equals:

$$\mathbf{\Sigma}_{11i} = \mathbf{\Lambda}_a\mathbf{\Lambda}_a^t + \mathbf{\Lambda}_e\mathbf{\Lambda}_e^t + \mathbf{\Lambda}_q\mathbf{\Lambda}_q^t + \mathbf{\Lambda}_d\mathbf{\Lambda}_d^t \qquad (7)$$

and the (pxp) cross covariance matrix $\mathbf{\Sigma}_{21i}$ equals:

$$\mathbf{\Sigma}_{21i} = \mathbf{\Lambda}_a\,[.5 \otimes \mathbf{I}]\,\mathbf{\Lambda}_a + \mathbf{\Lambda}_q\,[\pi_i \otimes \mathbf{I}]\,\mathbf{\Lambda}_q + \mathbf{\Lambda}_d\,[p[1]_i \otimes \mathbf{I}]\,\mathbf{\Lambda}_d, \qquad (8)$$

where $\otimes$ is kronecker matrix multiplication and **I** is the identity matrix of appropriate dimension (the result of $[.5 \otimes \mathbf{I}]$ is a diagonal matrix with .5 on the diagonal). Assuming the phenotypic data is approximately normally distributed, parameters in the matrices $\mathbf{\Lambda}_a$, $\mathbf{\Lambda}_q$, $\mathbf{\Lambda}_d$, and $\mathbf{\Lambda}_e$ can be estimated by maximizing the raw data loglikelihood function, and tests of significance based on the loglikelihood ratio test.

SIMPLIFYING ASSUMPTIONS AND MODEL PARAMETER VALUES

We assume that the QTL has 2 equi-frequent alleles and that its alleles are either codominant, or dominant. We assume that we have a marker situated zero cM away from the QTL, i.e. the QTL and the marker are adjacent on the chromosome. The marker has an infinite number of alleles (polymorphic information content, or PIC = 1), 16 alleles (PIC = .934), or 8 (PIC = .861). Regardless of the PIC value, the marker alleles are equi-frequent.

Under these simplifying assumptions, Table 17.2 in Haseman and Elston[9] can be used to derive a limited number of expected groups which are defined by different combinations of the values for π_i and $p[1]_i$. The number of sib-pairs within each group depends directly on the number of equi-frequent alleles at the marker locus, or, equivalently on the PIC value of the marker[20]. Depending on whether the

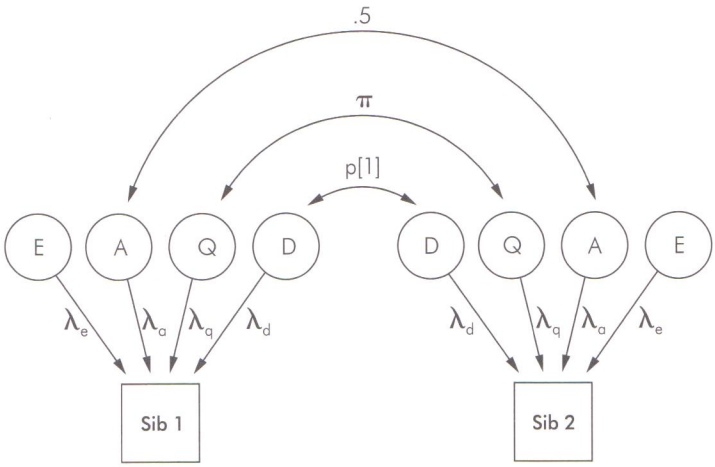

Figure 17.1 – Path diagram with observed phenotypes in sib 1 and sib 2 represented by squares and latent variables E (individual-specific environment), A (additive genetic background), Q (additive QTL effects) and D (non-additive QTL effects) represented by circles. The correlation between additive genetic background influences is 0.5, the correlation between additive QTL effects equals the proportion of alleles shared identical-by-descent and the correlation between non-additive QTL effects is p[1]: the probability that siblings share all alleles identical-by-descent.

number of markers, m, is infinite or not, and depending on whether the QTL is codominant or dominant, Table 17.1 shows that we have 3 (when $m=\infty$), 5 (when PIC < 1 and there is no dominance), or 7 distinct groups (PIC < 1 and dominance at the QTL locus). Because π_i and $p[1]_i$ only assume the 7 combinations shown in Table 17.1, we can use multi-group covariance structure modeling to estimate parameters.

For the power calculations we considered a scenario in which the QTL effect explains 25% of the total variance in a trait. Background genetic influences also account for 25% of the variance so that the total heritability of the trait is 50% and the amount of variance explained by (random) environmental factors is also 50% (actual values for variances used to construct the simulated covariance matrices were $Vq = \lambda_q^2 = 3$ (for the non-additive QTL $Vq = \lambda_q^2 = 2$ and $Vd = \lambda_d^2 = 1$), $Va = \lambda_a^2 = 3$ and $Ve = \lambda_e^2 = 6$. In the univariate covariance structure model we have: $\lambda_a = \sqrt{3}$, $\lambda_e = \sqrt{6}$, and $\lambda_q = \sqrt{3}$, and $\lambda_d = 0$, for the codominant QTL, or $\lambda_q = \sqrt{2}$, and $\lambda_d = \sqrt{1}$, for the dominant QTL.

In the multivariate case, we have 4 phenotypes. The interrelationship between

these phenotypes are determined by the matrices of regression coefficients. We introduce the following values in the case of a codominant QTL:

$$\mathbf{\Lambda}^t_a = [\sqrt{3} \quad \sqrt{3} \quad \sqrt{3} \quad \sqrt{3}]$$

$$\mathbf{\Lambda}_e = \begin{bmatrix} \sqrt{2} & \sqrt{4} & 0 & 0 & 0 \\ \sqrt{2} & 0 & \sqrt{4} & 0 & 0 \\ \sqrt{2} & 0 & 0 & \sqrt{4} & 0 \\ \sqrt{2} & 0 & 0 & 0 & \sqrt{4} \end{bmatrix}$$

$$\mathbf{\Lambda}^t_q = [\sqrt{3} \quad \sqrt{3} \quad \sqrt{3} \quad \sqrt{3}]$$

In the case of a dominant QTL:

$$\mathbf{\Lambda}^t_q = [\sqrt{2} \quad \sqrt{2} \quad \sqrt{2} \quad \sqrt{2}],$$

and

$$\mathbf{\Lambda}^t_d = [\sqrt{1} \quad \sqrt{1} \quad \sqrt{1} \quad \sqrt{1}].$$

In the multivariate case, we assume that the 4 variables are influences by a single additive polygenic factor, and a single QTL. The unshared environmental influences are in part common to the 4 phenotypes, and in part specific to the 4 phenotypes. As mentioned in the Introduction, the model may arise when a phenotype is measured repeatedly over a short time span.

POWER CALCULATIONS

We refer to the true model, including the QTL factor, as H1, and we refer to the false model, excluding the QTL, as H0. Power equals 1-β, where β is:

$$\beta = \text{prob(accepting H0} \mid \text{H1 is true)},$$

i.e., the probability of a type II error. In the present context this means that there is a QTL effect, but that it is not detected. To calculate the power we follow the procedure described in Neale and Cardon[2] (1992, p. 190 ff.; see also 3.22). First we construct covariance matrices according to the true model, which includes the QTL. Next we use the Mx program[4] to fit the false model to these matrices using maximum likelihood (ML) estimation. The total number of sib-pairs, N, is chosen arbitrarily in fitting this model (say, 1000, or 5000). As is clear from Table 17.1, the distribution of this total N over the groups depends on the number of equi-frequent marker alleles and on the gene action of the QTL (codominant or dominant). The distribution of the goodness of fit index under the false model is distributed as a non-central chi-square variate. The exact form of the distribution

Table 17.1 – Distribution of π and p[1] given m, number of equi-frequent marker alleles, for m=8 (PIC=.861), m=16 (PIC=.934), m=32 (PIC=.968), and m= ∞ (PIC=1)*.

group	π	p[1]	frequency	m=8	m=16	m=32	m=[∞]
1	0	0	$1/4(m^3-2m^2+1)/m^3$	.1879	.2188	.2343	0.25
2	0.25	0	$(m-1)/m^2$	.1093	.0585	.0302	0
3	0.50	0	$1/2(m^2-2m+1)/m^2$	.3828	.4394	.4692	0.5
4	0.50	0.25	$1/m^2$	.0156	.0039	.0009	0
5	0.50	0.50	$1/2(m-1)/m^3$	.0068	.0018	.0004	0
6	0.75	0.50	$(m-1)/m^2$	.1093	.0585	.0302	0
7	1	1	$1/4(m^3-2m^2+1)/m^3$	.1879	.2188	.2343	0.25

* In the event of PIC = 1, we have three groups (1,3,7); in the event of PIC<1 and a codominant QTL, we have 5 groups (groups 3,4,5,are collapsed into a single group); in the event of PIC < 1 and a dominant (or recessive) QTL, we have 7 groups.

depends on the number of degrees of freedom, and the so-called non-centrality parameter (NCP). The number of degrees of freedom is simply the difference in the number of parameters between the true model (including the QTL) and the false model (without the QTL). The NCP equals the chi-square for the false model as reported by the program (i.e., Mx). Given N, the non-centrality parameter and the pre-specified α (e.g., .05, or .001), one can calculate the power to reject the false model, and one can calculate the required N to reject the false model, given a predetermined power. Conveniently, Mx carries out all the necessary calculations automatically. Below we report the required number of sib-pairs to reject the false hypothesis, given a power of .80 and an α of .001.

RESULTS AND DISCUSSION

Table 17.2 and 17.3 contain the number of sibling pairs needed to detect the effect of a QTL explaining 25% of the phenotypic variance. Table 17.2 summarizes the power calculations for a codominant QTL. As is expected (Fulker and Cherny, 1996), fitting the bivariate model gives better results in terms of power than analyzing squared difference scores. Regardless of PIC, the latter requires about a factor 1.35 more subjects than the former to achieve the same power. Multivariate model fitting involving all 4 phenotypes gives a substantial increase in power: 65% fewer subjects are required than in the univariate analysis. If the loadings of the QTL on the repeated measures of the phenotype can be constrained equal to each other, the increase in power is even larger, because the QTL effect can then be tested against 1 degree of freedom. The effect of PIC is as expected: the more informative the marker is, the more powerful the test of the QTL. Regardless of the test used, the reduction in the number of required sibpairs is about the same (from PIC=.93 to 1.0, about a factor .93). In terms of an ANOVA, one could say

that PIC and 'type of test' both have a main effect on the required number of sibpairs, but that an interaction is absent.

Table 17.3 presents the number of sibling pairs required to detect the presence of a dominant QTL effect, the additive QTL component and the test of a QTL effect when dominance is ignored when fitting the full model (last 4 rows in Table 17.3). First, it is clear that the detection of the dominance variance of the QTL requires very large sample sizes. Multivariate modeling does substantially reduce the number of required sibpairs, but even the most powerful test still required over 16000 sibpairs. The power to detect the presence of the additive and dominance QTL variance simultaneously is much greater (second 4 rows in Table 17.3). Here the required samples sizes are comparable to those shown in Table 17.2. As it is very difficult to detect the dominance deviation, we finally investigate the power to detect the dominant QTL, under the circumstance that it is fit as a codominant QTL. This means that we model the QTL effect using a single parameter. The result (last 4 rows in Table 17.3) are very similar to those shown in Table 17.2. As in Table 17.2, there does not seem to be any interaction between the effects of PIC and the effects of 'type of test'.

The considerable increase in power associated with the multivariate test, suggest that it is advisable to collect multiple indicators of the phenotype under consideration or measure the phenotype repeatedly at multiple timepoints. An interesting question concerns the returns in terms of power of adding indicators. Figure 17.2 displays the required number of subjects to detect the codominant QTL when 1 to 9 indicators are analyzed. Again we consider the same three PIC values.

In Figure 17.2 we see that there is a dramatic increase in power when going from 1 to 2 and from 2 to 3 indicators. Beyond 3 indicators the increase in power is small, and beyond 5 indicators, the power actually decreases. Although the minimum number of required subjects is observed at 5 indicators, 3 or 4 indicators are sufficient. Needless to say, these particular results cannot be generalized to other parameter values, or genetic covariance structure models. However, it is very likely that the observed diminished returns will hold regardless of the details relating to the model.

In an earlier paper[8] we explored several strategies to analyze multivariate phenotypes. We found that when the multivariate information was summarized into a genetic factor score[23,24] no information was lost compared to fitting the complete multivariate model. This is a useful result because working with multivariate phenotypes may pose a problem in studies that selectively genotype extreme scoring sibling pairs. Multivariate selection of such pairs can be carried out on a genetic factor score which represents a subjects score on the latent genetic factor underlying the observations.

There are several ways to include a QTL in GCSM, which can be denoted the pi-

Table 17.2 – Number of sib-pairs to detect a codominant QTL with power=.80 and α=0.001. The QTL accounts for 25%, background genes for 25% and environment for 50% of the total variance. For the multivariate data these effect sizes are the same for all 4 variables; environmental influences are split into variable specific effects (33%) and a common factor effect (17%).

analysis	PIC=1	PIC=.93	PIC=.86
Squared Difference score (df = 1)	2434	2598	2819
SEM Univariate (df = 1)	1795	1915	2077
SEM Multivariate (df = 4)	1155	1234	1340
SEM Multivariate (df = 1)	854	912	990

Table 17.3 – Number of sib-pairs to detect a QTL with equal allele frequencies, QTL additive effect = 16.6%, QTL non-additive effect = 8.3% (other effect sizes as in table 17.2; power=.80 and α=0.001).

	PIC=1	**PIC=.93**	**PIC=.86**
Dominance effect			
Difference score (df = 1)	47,534	53,755	61,477
SEM Univariate (df = 1)	33,039	37,442	42,947
SEM Multivariate (df = 4)	22,052	24,928	28,522
SEM Multivariate (df = 1)	16,300	18,426	21,083
Dominance+Additive effect			
Difference score (df = 2)	2664	2861	3123
SEM Univariate (df = 2)	1961	2103	2291
SEM Multivariate (df = 8)	1324	1423	1554
SEM Multivariate (df = 2)	936	1005	1098
Total (D+A) effect			
Difference score (df = 1)	2432	2605	2836
SEM Univariate (df = 1)	1795	1920	2086
SEM Multivariate (df = 4)	1157	1240	1350
SEM Multivariate (df = 1)	855	916	998

hat approach and the IBD-distribution, or mixture, approach[6,7,25]. As it was more convenient for our present purposes, we have used the pi-hat approach in our power calculations. In unselected samples, these two approaches produce almost identical results.

In conclusion, on the basis of the present results, it appears that GCSM has more power than the original Haseman and Elston regression method[7] and that multivariate GCSM is more powerful than univariate GCSM[20].

that PIC and 'type of test' both have a main effect on the required number of sibpairs, but that an interaction is absent.

Table 17.3 presents the number of sibling pairs required to detect the presence of a dominant QTL effect, the additive QTL component and the test of a QTL effect when dominance is ignored when fitting the full model (last 4 rows in Table 17.3). First, it is clear that the detection of the dominance variance of the QTL requires very large sample sizes. Multivariate modeling does substantially reduce the number of required sibpairs, but even the most powerful test still required over 16000 sibpairs. The power to detect the presence of the additive and dominance QTL variance simultaneously is much greater (second 4 rows in Table 17.3). Here the required samples sizes are comparable to those shown in Table 17.2. As it is very difficult to detect the dominance deviation, we finally investigate the power to detect the dominant QTL, under the circumstance that it is fit as a codominant QTL. This means that we model the QTL effect using a single parameter. The result (last 4 rows in Table 17.3) are very similar to those shown in Table 17.2. As in Table 17.2, there does not seem to be any interaction between the effects of PIC and the effects of 'type of test'.

The considerable increase in power associated with the multivariate test, suggest that it is advisable to collect multiple indicators of the phenotype under consideration or measure the phenotype repeatedly at multiple timepoints. An interesting question concerns the returns in terms of power of adding indicators. Figure 17.2 displays the required number of subjects to detect the codominant QTL when 1 to 9 indicators are analyzed. Again we consider the same three PIC values.

In Figure 17.2 we see that there is a dramatic increase in power when going from 1 to 2 and from 2 to 3 indicators. Beyond 3 indicators the increase in power is small, and beyond 5 indicators, the power actually decreases. Although the minimum number of required subjects is observed at 5 indicators, 3 or 4 indicators are sufficient. Needless to say, these particular results cannot be generalized to other parameter values, or genetic covariance structure models. However, it is very likely that the observed diminished returns will hold regardless of the details relating to the model.

In an earlier paper[8] we explored several strategies to analyze multivariate phenotypes. We found that when the multivariate information was summarized into a genetic factor score[23,24] no information was lost compared to fitting the complete multivariate model. This is a useful result because working with multivariate phenotypes may pose a problem in studies that selectively genotype extreme scoring sibling pairs. Multivariate selection of such pairs can be carried out on a genetic factor score which represents a subjects score on the latent genetic factor underlying the observations.

There are several ways to include a QTL in GCSM, which can be denoted the pi-

Table 17.2 – Number of sib-pairs to detect a codominant QTL with power=.80 and α=0.001. The QTL accounts for 25%, background genes for 25% and environment for 50% of the total variance. For the multivariate data these effect sizes are the same for all 4 variables; environmental influences are split into variable specific effects (33%) and a common factor effect (17%).

analysis	PIC=1	PIC=.93	PIC=.86
Squared Difference score (df = 1)	2434	2598	2819
SEM Univariate (df = 1)	1795	1915	2077
SEM Multivariate (df = 4)	1155	1234	1340
SEM Multivariate (df = 1)	854	912	990

Table 17.3 – Number of sib-pairs to detect a QTL with equal allele frequencies, QTL additive effect = 16.6%, QTL non-additive effect = 8.3% (other effect sizes as in table 17.2; power=.80 and α=0.001).

	PIC=1	**PIC=.93**	**PIC=.86**
Dominance effect			
Difference score (df = 1)	47,534	53,755	61,477
SEM Univariate (df = 1)	33,039	37,442	42,947
SEM Multivariate (df = 4)	22,052	24,928	28,522
SEM Multivariate (df = 1)	16,300	18,426	21,083
Dominance+Additive effect			
Difference score (df = 2)	2664	2861	3123
SEM Univariate (df = 2)	1961	2103	2291
SEM Multivariate (df = 8)	1324	1423	1554
SEM Multivariate (df = 2)	936	1005	1098
Total (D+A) effect			
Difference score (df = 1)	2432	2605	2836
SEM Univariate (df = 1)	1795	1920	2086
SEM Multivariate (df = 4)	1157	1240	1350
SEM Multivariate (df = 1)	855	916	998

hat approach and the IBD-distribution, or mixture, approach[6,7,25]. As it was more convenient for our present purposes, we have used the pi-hat approach in our power calculations. In unselected samples, these two approaches produce almost identical results.

In conclusion, on the basis of the present results, it appears that GCSM has more power than the original Haseman and Elston regression method[7] and that multivariate GCSM is more powerful than univariate GCSM[20].

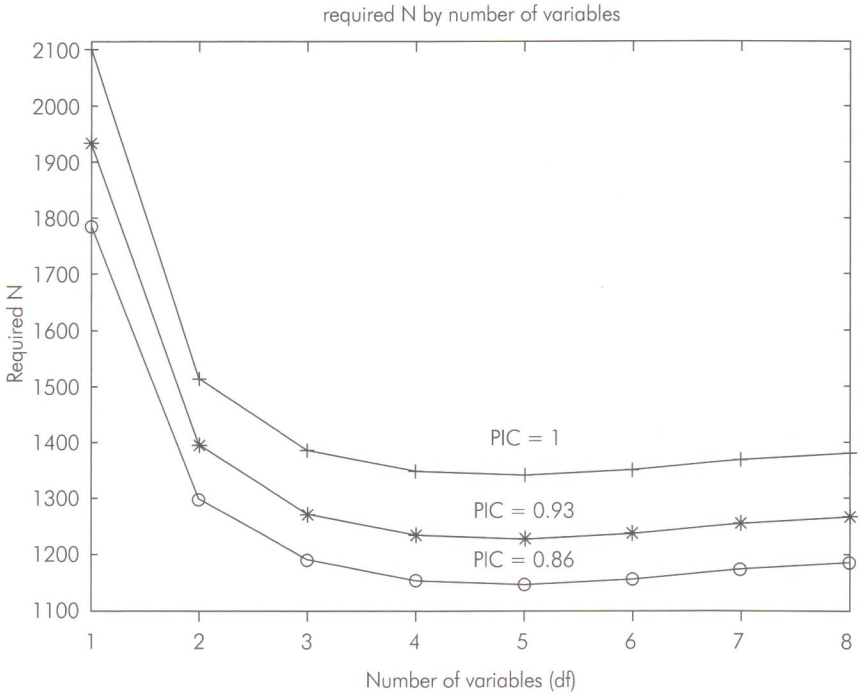

Figure 17.2 – number of sibpairs required to attain α=.001 and 1-β=.80 as a function of the number of phenotypic indicators (parameter values are the same as those in Table 17.1). The three plots (top to bottom) correspond to PIC=1, PIC=.93, and PIC=.86, respectively.

ACKNOWLEDGEMENT

We thank Mike Neale for helping with the model specification of a dominant QTL.

REFERENCES

1 Martin N.G. and Eaves L.J. (1977). The genetical analysis of covariance structure. *Heredity*, *38*, 79–95, 1977.

2 Neale M.C. and Cardon L.R.(1992). *Methodology for Genetic Studies of Twins and Families* (NATO ASI Series D: Behavioural and Social Sciences-Vol. 67), Dordrecht: Kluwer Academic Publishers.

3 Jöreskog, K.G. and Sörbom, D. (1989). *LISREL 7: A guide to the program and applications* (2nd.). Chicago: SPSS press.

4 Neale M. (1997). *Statistical Modeling with Mx*, Department of Human Genetics, Box 3, MCV, Richmond VA 23298.

5 Boomsma D.I. and Molenaar P.C.M. (1986). Using LISREL to analyze genetic and environmental covariance structure, *Behavior Genetics*, *16*, 237–250.

6 Eaves L.J., Neale M. and Maes H. (1996). Multivariate multipoint linkage analysis of quantitative trait loci. *Behavior Genetics*, *26*, 519–525.

7 Fulker D.W., Cherny S.S., and Cardon L.R. (1995). Multipoint interval mapping of quantitative trait loci, using sib pairs. *American Journal of Human Genetics*, *56*, 1224–1233.

8 Martin N., Boomsma D., and Machin, G. (1997). A twin-pronged attack on complex traits. *Nature Genetics*, *17*, 387–391.

9 Haseman J.K. and Elston R.C. (1972). The investigation of linkage between a quantitative trait and a marker locus. *Behavior Genetics*, *2*, 3–19.

10 Goldgar D.E. (1990). Multipoint analysis of human quantitative genetic variation. *American Journal of Human Genetics*, *47*, 957–967.

11 Fulker D.W. and Cherny S.S. (1996). An improved multipoint sib-pair analysis of quantitative traits. *Behavior Genetics*, 26, 527–532.

12 Kruglyak L. and Lander E. (1995). Complete multipoint sib pair analysis of qualitative and quantitative traits. *American Journal of Human Genetics*,, 57, 439–454.

13 Carey G. and Williamson J.(1991). Linkage analysis of quantitative traits: Increased power by using selected samples. *American Journal Human Genetics*, *49*, 786–796, 1991.

14 Eaves L. and Meyer J. (1994). Locating human quantitative trait loci: Guidelines for the selection of sibling pairs for genotyping. *Behavior Genetics*, *24*, 443–455.

15 Cardon L.R. and Fulker D.W. (1994). The power of interval mapping of quantitative trait loci using selected sib pairs. *American Journal of Human Genetics*, *55*, 825–833.

16 Risch N. and Zhang H. (1995). Extreme discordant sib pairs for mapping quantitative trait loci in humans, *Science*, *268*, 1584–1589.

17 Gu C., Todorov A., and Rao D.C.(1996). Combining extremely concordant sibpairs with extremely disconcordant sibpairs provides a cost effective way to linkage analysis of quantitative trait loci, *Genetic Epidemiology*, 13, 513–533.

18 Dolan C.V. and Boomsma D.I. (1998). Optimal selection of sib-pairs from random samples for linkage analysis of a QTL using the EDAC test. *Behavior Genetics*, 28, 197–206.

19 Matsueda R.L. and Bielby W.T. (1986). Statistical power in covariance structure models. In: N. B. Tuma (Ed.). *Sociological Methodology, 1986*. San Fransisco: Jossey-Bass.

20 Boomsma D.I. and Dolan C.V. (1998). A comparison of power to detect a QTL in sib-pair data using multivariate phenotypes, mean phenotypes, and factor-scores, Behavior Genetics, 28, 329–340.

21 Sham P. (1998). *Statistics in Human Genetics*. New York: John Wiley & Sons.

22 Saris, W.E. and Satorra, A. (1993). Power Evaluations in Structural Equation Modeling. In: K.A. Bollen and J. S. Long, (Eds.). *Testing structural equation models*. p.181–204. Newbury Park: Sage Publications.

23 Boomsma D.I., Molenaar P.C.M., Orlebeke J.F.(1990). Estimation of individual genetic and environmental factor scores, *Genetic Epidemiology*, 7, 83–91.

24 Boomsma D.I., Molenaar P.C.M., Dolan C.V. (1991). Estimation of individual genetic and environmental profiles in longitudinal designs, *Behavior Genetics*, 21, 241–253.

25 Dolan C.V., Boomsma D.I., Neale MC, A simulation study of the effects of assigning prior IBD probabilities to unselected sib-pairs in covariance structure modeling of a QTL test, Am J Human Genetics, 64, 268–280, 1999.