Behavior Genetics, Vol. 26, No. 2, 1996

Using Multivariate Genetic Modeling to Detect Pleiotropic Quantitative Trait Loci

Dorret I. Boomsma[1]

Received 3 July 1995—Final 20 Oct. 1995

Large numbers of sibling pairs or other relatives are needed to detect linkage between a quantitative trait locus (QTL) and a marker, especially if the variance of the QTL is low relative to the total phenotypic variance of the trait. One strategy to increase the power to detect linkage is to reduce the environmental variance in the trait under analysis. This approach was explored by carrying out a series of simulation studies in which multivariate observations were used to estimate individual genotypic values at a QTL, that pleiotropically affected more than one trait. Simulations for different QTL allele frequencies with a completely informative marker showed that the power to detect the QTL increased substantially when estimates of individual genotypic values at the QTL were used in the linkage analysis instead of phenotypic observations. An advantage of this approach is that, rather than employing phenotypic selection, individuals with extreme genotypes may be selected when ascertaining a sample of extreme families.

KEY WORDS: Genotypic factor scores; multivariate genetic modeling; linkage analysis; quantitative trait locus.

INTRODUCTION

With the availability of highly polymorphic markers the genetic mapping of behavioral traits in humans has become possible. Several strategies have been developed to map quantitative trait loci (QTL), which are based on identifying marker alleles that are inherited identical by descent (IBD). Robust methods that study the genetic linkage of a quantitative trait and a polymorphic marker in data from sibling pairs have been developed, for example by Penrose (1938) and Haseman and Elston (1972). Amos and Elston (1989) have extended these methods to other types of noninbred relative pairs. These methods suppose that if a marker is cosegregating with a quantitative trait, then siblings whose trait values are more alike are more likely to receive the same alleles identical by descent at a closely linked marker locus than siblings whose resemblance for the trait is less.

However, even with large numbers of highly polymorphic markers that allow the IBD status of family members such as siblings to be known, the power to detect a single locus that influences quantitative traits in humans is low (e.g., Blackwelder and Elston, 1982). One strategy to increase power to detect genetic linkage is to reduce the environmental variance of the phenotype under study. The structural equation model that can be used to analyze multivariate data from genetically informative samples offers the possibility to statistically reduce environmental variance by estimating an individuals genotypic value at a QTL. This approach was investigated by carrying out a series of simulations in which individual genotypic scores were first estimated from multivariate phenotypes, which were all influenced by the same quantitative trait locus, and then used in a linkage analysis.

[1] Department of Psychonomics, Vrije Universiteit, De Boelelaan 1111, 1081 HV Amsterdam, The Netherlands. Fax: 31-20-4448832; e-mail: dorret@psy.vu.nl.

0001-8244/96/0300-0161$09.50/0

Individual genotypic scores may be estimated from multivariate data measured in genetically informative relatives. Multivariate observations from family members can be used to test whether the same genetic factor pleiotropically influences multiple phenotypically correlated measures (Martin and Eaves, 1977; Boomsma and Molenaar, 1986). If a common genetic factor is found (see Fig. 1), scores on this factor can be constructed for an individual by standard methods for the estimation of factor scores (Boomsma *et al.*, 1990, 1991). Factor scores cannot be estimated in the usual statistical sense, since they are not parameter values but values ascribed to unobservable variates belonging to an individual (Lawley and Maxwell, 1971, Chap. 8). Because the number of observations usually is smaller than the number of latent factors, it is necessary to introduce a minimum variance or least squares principle to estimate individual factor scores [see Sans *et al.* (1978) for a review of different methods]. The regression method for the estimation of factor scores minimizes the sum of squares of the difference between estimated and true factor scores and is the preferred method when the primary interest is in the factor scores themselves.

We have shown that the regression method may be successfully applied in multivariate genetic modeling and that in both cross-sectional and longitudinal designs, individual estimates of factor scores can be reliably obtained (Boomsma *et al.*, 1990, 1991).

In this article the regression method to estimate factor scores was applied to simulated multivariate twin data to address the question whether power to detect a QTL could be increased by using individual genotypic factor scores in a linkage analysis. Simulations were carried out to compare the power of the Haseman and Elston (1972) regression method for linkage analysis with observed quantitative phenotypes and with estimated individual genotypic scores. Three-variate phenotypes were simulated for MZ and DZ twin pairs. The total heritability of all three phenotypes was .5 and the heritability of the QTL was .25. Data from MZ and DZ twins were used to fit the multivariate model to the phenotypic observations. Data from DZ twins were used in the linkage analysis. By taking this approach, not only is the environmental variance in the data accounted for, but also the background genetic variation that is not associated with the QTL.

SIMULATIONS

(1) QTL and marker data were generated for 1000 fathers and 1000 mothers. Only heterozygous parents were simulated and heterozygotes mated only with heterozygotes that carried different alleles than they carried themselves, so that the parents had four marker alleles. Thus IBD status of their offspring was always known for certain.

(2) The QTL had two alleles in Hardy–Weinberg equilibrium that were not associated with the marker alleles. A represents the increaser allele at the QTL with frequency p, the genotypic value of AA is d; a represents the decreaser allele with frequency $q = 1 - p$ and genotypic value $-d$. There was no dominance, the genotypic value of Aa was 0. Three allele frequencies for the increaser allele were considered: .5, .7, and .9. Given these allele frequencies and an additive genetic variance at the QTL of 1, the value of d was obtained by solving for the variance of the QTL: $\sigma^2 = 1 = 2pqd^2$. This gives $d = 1.543$ for $p = .7$ (with QTL mean $\mu = (p - q)d = .617$), and $d = 1.414$ and 2.357 for allele frequencies .5 and .9, respectively.

(3) For DZ twin pairs two parents were drawn without replacement from the parental population and one of their chromosomes was randomly selected for each of their two children. This gave QTL values for each sibling and IBD status (0,1,2) for sibling pairs.

(4) Three phenotypes were created for each subject according to (see also Fig. 1):

$$P(ij) = \lambda_q(i)*\mathrm{QTL}(j) + \lambda_e(i)*E(j) + G(ij) + U(ij)$$

where each phenotype P ($i = 1, 2, 3$) for each subject ($j = 1, \ldots, N$) is a function of the QTL, which influences all three phenotypes and a function of an environmental factor (E) that is uncorrelated in family members, but also influences all three phenotypes. λ_q and λ_e are factor loadings of the phenotype on the QTL and the environmental factor common to all three phenotypes. Each phenotype is also influenced by a unique genetic factor (G) and a unique environmental factor (U). Within sibling pairs the QTL and the unique genetic factors are correlated .5 on average. The variance of all latent factors was 1, and all factor loadings λ_q

and λ_e also equaled 1. The variance of each of the three phenotypes thus was equal to 4, and their heritability to 2/4 = .5. The heritability of the QTL was 1/4 = .25.

(5) The same model was used to simulate data for MZ twins, who always have the same QTL values and the same unique genetic scores (*G*). No marker data were generated for MZ twins; their data were used only to estimate the genetic and environmental factor loadings in the multivariate model, so that individual factor scores could be estimated.

(6) Data sets for MZ and DZ pairs were created. Simulations with 200, 400, and 600 pairs of MZ and DZ twins were considered. One thousand replications were simulated for allele frequencies p = .5, .7, and .9, without recombination between the marker and the QTL. For allele frequency p = .5, two additional series of 1000 simulations were carried out, with recombination fractions Θ = .05 and Θ = .1.

ANALYSES

(1) The true multivariate model was fitted to the simulated data and factor loadings on the common QTL and *E* factors and the unique variances of *G* and *U* were estimated. These estimates were used for the construction of a weight matrix, according to the regression method, to obtain individual factor scores. The weight matrix was used to compute factor scores for each subject, using both his own multivariate phenotypic data and the data from his sibling:

$$f = A'P$$

where

f = [QTL1, QTL2, E1, E2] is a vector of factor scores of sib1 and sib2 to be estimated,
P = the measured multivariate phenotype of observations in sib1 and sib2,
A = weight matrix that is constant across subjects and depends on factor loadings and unique variances.

A is obtained by minimizing the sum of squares of the difference between estimated and true factor scores:

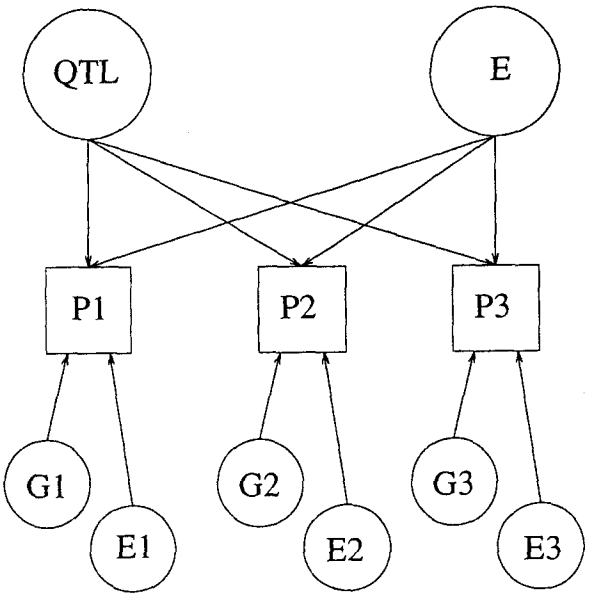

Fig. 1. Path model showing quantitative trait locus (QTL) and environmental factors common to three phenotypes plus unique genetic (*G*) and environmental (*U*) factors associated with each trait. Because the QTL pleiotropically influences more than one phenotype, estimates of individual QTL values may be obtained by standard methods for the estimation of factor scores.

$$A = \Psi\Lambda'\Sigma^{-1} = \Psi\Lambda'(\Lambda\Psi\Lambda' + H)^{-1}$$

where

Λ = matrix of loadings on QTL and *E* factors
Ψ = correlation matrix of factor scores
H = matrix of unique genetic and environmental variances

(2) The three phenotypic observations and the estimated genotypic factor scores were used in a linkage analysis. The Haseman and Elston (1972) sib-pair approach for linkage analysis was employed in which the squared difference between the scores of siblings (either their estimated QTL values or their observed phenotypic values) is regressed on the proportion of alleles shared IBD at the marker (π = IBD/2):

$$Y = \alpha + \beta\pi$$

where Y is the squared difference between quantitative trait values of siblings and π is the proportion of alleles IBD at the marker (π = 0, .5, 1). Haseman and Elston (1972) showed that $\alpha = \sigma_e^2 +$

Table I. Power to Detect Linkage Between a Two-Allele Quantitative Trait Locus (QTL) and a Fully Informative Polymorphic Marker Using the Haseman–Elston Regression Approach (with $\alpha = .05$) for Different Numbers of Sib Pairs[a]

Allele frequency and recombination fraction	Percentage					
	$N = 200$ pairs		$N = 400$ pairs		$N = 600$ pairs	
	Estimated QTL	Observed phenotype	Estimated QTL	Observed phenotype	Estimated QTL	Observed phenotype
$p = .5$ $\Theta = 0$	43	21	72	38	88	53
$p = .7$ $\Theta = 0$	40	21	70	39	88	53
$p = .9$ $\Theta = 0$	36	18	69	38	84	51
$p = .5$ $\Theta = .05$	28	13	49	25	68	36
$p = .5$ $\Theta = .10$	18	10	33	18	47	24

[a] Results based on 1000 simulations for each allele frequency (p) and recombination fraction (Θ).

$2\sigma_g^2$ and $\beta = -2(1 - 2\Theta)^2\sigma_g^2$, where Θ is the recombination fraction between the QTL and the marker and σ_g^2 is the additive genetic variance of the QTL. If the regression is negative and significant, it implies linkage with either a large QTL at some distance from the marker or a smaller QTL closer to the marker locus.

RESULTS

Table I first gives the outcomes of the linkage analyses for the cases in which there is no recombination between the QTL and the marker. Using the phenotypic observations in the analysis shows the well-known low power to detect linkage with a quantitative phenotype. Even when 600 pairs of siblings are available, the power is only around 50%. In contrast, for all allele frequencies, the power is increased substantially when estimated genotypic values are used in the analysis instead of observed phenotypic values. When there is recombination between the QTL and the marker, the increase in power still remains roughly twice as high for genotypic factor scores compared to phenotypic observations.

DISCUSSION

The results presented in this paper demonstrate that power to detect linkage in a sibling analysis of quantitative traits can be increased substantially by analyzing unobserved, estimated, genotypes instead of observed, measured, phenotypes. Using the information contained in the covariance between quantitative traits in this way leads to a substantial increase in power to detect quantitative trait loci.

The method of estimating individual scores on latent factors is well established in factor analysis (Lawley and Maxwell, 1971; Mulaik, 1972). To be applied in genetic modeling, the method requires multiple QTL indicators measured in genetically related individuals. As more or better indicators of the QTL are available, individual factor scores can be obtained more reliably. Estimation of individual genotypic and environmental scores is numerically also possible in univariate designs, but this gives intercorrelated estimates of independent factor scores. In a univariate design, for example, DZ twins supply 2 observations (one on twin1 and one on twin2). Even under a simple additive genetic model, this does not provide enough information to obtain independent factor scores, since four factor scores need to be estimated (two genotypic scores that are correlated in siblings and two environmental scores that are uncorrelated). The method is not restricted to twin data and can easily be generalized to multivariate data from other family members.

In their 1989 paper Lander and Botstein suggested three strategies to increase power in QTL mapping. The first two methods, selective genotyping of extreme phenotypes and interval mapping or

simultaneous search, have been explored in several papers. Carey and Williamson (1991), Cardon and Fulker (1994), Eaves and Meyer (1994), and Risch and Zhang (1995) have extensively discussed the value of nonrandom sampling strategies and shown that selective genotyping of extreme individuals may lead to an appreciable difference in power of linkage studies of quantitative traits.

Likewise, methods for QTL multipoint interval mapping in humans have successfully been developed. Goldgar (1990) introduced a method for estimating the proportion of genetic material IBD in a chromosomal region based on marker loci spanning the region and incorporated these estimates into a variance-components model. Comparing this approach to the Haseman and Elston method showed it to be more powerful. Fulker *et al.* (1994, 1995) demonstrated both an increase in power and good prospects for approximate QTL location employing interval mapping methods based on multiple regression. Recently, Kruglyak and Lander (1995) described how to obtain the complete multipoint inheritance information for sibling pairs, which uses all available marker information, and how to employ this information to map both qualitative and quantitative traits.

The approach outlined in this paper represents an example of the third strategy to increase power in QTL mapping, i.e., statistically decreasing environmental variance by using an estimate of an individual's genotypic value at a QTL. This approach also decreases the background genetic variance that is not associated with the QTL. The three approaches to increase power, selective genotyping, multipoint mapping, and reduction of environmental and genetic background variation are not mutually exclusive and may, in fact, be employed simultaneously. The largest increases in power probably will be realized by combining them into one design.

It remains to be established how the approach outlined in this paper relates to fitting the complete multivariate model to the data simultaneously with the marker information. Since factor scores can be estimated only imprecisely, it is likely that fitting the complete model to the data will show even larger increases in power than the approach outlined in this paper. Amos *et al.* (1990) explored how multiple measures can be incorporated into a multivariate regression approach that estimates the linear function that results in the strongest correlation between the squared pair differences and IBD status at the marker locus. Schork (1993) applied the method proposed by Goldgar (1990) to bivariate phenotypes and found that a larger genetic correlation between the two traits led to a larger increase in power. However, these last methods always require marker information on all subjects.

One clear advantage of the method of estimated genotypic factor scores compared to general multivariate models is that selective genotyping of extreme individuals can be based on sibling pairs that have been selected on the basis of their extreme genotype, instead of their phenotype.

REFERENCES

Amos, C. I., and Elston, R. C. (1989). Robust methods for the detection of genetic linkage for quantitative data from pedigrees. *Genet. Epidemiol.* **6**:349–360.

Amos, C. I., Elston, R. C., Bonney, G. E., Keats, B. J. B., and Berenson, G. S. (1990). A multivariate method for detecting genetic linkage, with application to a pedigree with an adverse lipoprotein phenotype. *Am. J. Hum. Genet.* **47**:247–254.

Blackwelder, W. C., and Elston, R. C. (1982). Power and robustness of sib-pair linkage tests and extension to larger sibships. *Comm. Stat. Theory Meth.* **11**:449–484.

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

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

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

Cardon, L. R., and Fulker, D. W. (1994). The power of interval mapping of quantitative trait loci using selected sib pairs. *Am. J. Hum. Genet.* **55**:825–833.

Carey, G., and Williamson, J. (1991). Linkage analysis of quantitative traits: Increased power by using selected samples. *Am. J. Hum. Genet.* **49**:786–796.

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

Fulker, D. W., and Cardon, L. R.(1994). A sib-pair approach to interval mapping of quantitative trait loci. *Am. J. Hum. Genet.* **54**:1092–1103.

Fulker, D. W., Cherney, S. S., and Cardon, L. R. (1995). Multipoint interval mapping of quantitative trait loci, using sib pairs. *Am. J. Hum. Genet.* **56**:1224–1233.

Goldgar, D. E. (1990). Multipoint analysis of human quantitative genetic variation. *Am. J. Hum. Genet.* **47**:957–967.

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

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

Lawley, D. N., and Maxwell, A. E. (1971). *Factor Analysis as a Statistical Method*, Butterworths, London.

Martin, N. G., and Eaves, L. J. (1977). The genetical analysis of covariance structure. *Heredity* **38**:79–95.
Mulaik, S. A. (1972). *The Foundations of Factor Analysis*, McGraw–Hill, New York.
Penrose, G. S. (1983). Genetic linkage in graded human characters. *Ann. Eugen.* **8**:223–237.
Risch, N., and Zhang, H. (1995). Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science* **268**:1584–1589.
Saris, W. E., De Pijper, M., and Mulder, J. (1978). Optimal procedures for estimation of factor scores. *Sociol. Methods Res.* 7:85–106.
Schork, N. J. (1993). Extended multipoint identity-by-descent analysis of human quantitative traits: Efficiency, power, and modeling considerations. *Am. J. Hum. Genet.* **53**: 1306–1319.

Edited by Kay Phillips